# Assessing activity energy expenditure from body-worn sensors during free-living


Hughes Hall

This dissertation is submitted for the degree of Doctor of Philosophy.

September 2018

Author:

Thomas White

Supervisor:

Dr. Søren Brage

# Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee (60,000 words).

# Thesis summary

**Title:**

Assessing activity energy expenditure from body-worn sensors during free-living

**Author:**

Thomas White

There has been widespread adoption of single body-worn sensors to objectively capture the physical activity of free-living individuals in large studies across the world. For research into metabolic diseases such as obesity and diabetes, it is useful to use this data to assess activity energy expenditure, which requires development of inference models.

This thesis describes the derivation and evaluation of models to estimate activity energy expenditure from acceleration data collected at either wrist or thigh. Two fundamentally different approaches were pursued; one follows a traditional approach of regressing metrics of movement intensity against activity energy expenditure, and one uses neural networks to learn a more complex relationship directly from the raw data. The performance of these models was then evaluated by agreement with a gold standard measure of energy expenditure in free-living humans. The generalisability of these models was then investigated by validating them in a large African cohort. Finally, the differences between the two methodological approaches were explored using a dataset of everyday

activities performed in a laboratory.

The movement intensity models accurately and precisely estimated activity energy expenditure in free-living adults with small and non-significant mean biases at the population level, and the neural network models offered a relatively modest but consistent increase in performance over their movement intensity counterparts. All models appeared to overestimate activity energy expenditure in the African population, which suggests that population specificity is a possibility, and caution should therefore be used when making international comparisons. There were systematic differences between the two modelling approaches when examined by activity type, indicating that the neural networks may be implicitly recognising activities, which may facilitate activity classification in free-living in the future. This works enhances the utility of raw acceleration signals now being collected in several large studies worldwide, and highlights the need for population-specific validity evaluation.

# Acknowledgements

I would like to thank my supervisor, Dr Soren Brage, for his excellent guidance and his boundless enthusiasm and energy. I can only hope for a lifetime of collaborations as stimulating and productive as ours has been.

My sincere gratitude to the Physical Activity Technical Team, particularly Kate Westgate, Stefanie Hollidge, and Lewis Griffiths, for their extraordinary work in conducting our doubly labelled water study, and for patiently answering an uncountable number of questions over the years.

Thank you to my external collaborators Dr Yu Guan and Dr Thomas Ploetz for their advice on the use of deep neural networks, which became such a significant part of this thesis.

My thanks to MedImmune for awarding me a studentship in order to conduct this research, and to both MedImmune and Newcastle University for covering the material costs of the doubly labelled water study.

On an informal note, I wish to thank all of my friends for regularly assisting me in the acquisition and maintenance of a calming state of adequate inebriation in the evenings. And thank you to the morning coffee crew at the MRC Epidemiology Unit, for giving me a strong incentive to get out of bed and into the office at a sensible time every day.

# Structure of this thesis

The following thesis is divided into seven chapters. The first chapter contains background introductory material, to familiarise the reader with standard practices in the processing of accelerometry data, and the ways in which free-living physical activity can be characterised using such signals.

Chapters 2 through to Chapter 6 are “analysis” chapters, each of which describe a self-contained piece of original work lead by the thesis author. These chapters are presented in the form of journal article manuscripts (at the time of writing, the first two analysis chapters are already in print, and two chapters are prepared for submission as separate articles). As a consequence, readers will notice that there is (by necessity) some overlap and redundancy in their content, particularly in the introduction and methods sections.

The final chapter concludes the thesis by summarising the findings of each chapter, describing how they fit together into a cohesive line of investigation, and reflecting on the body of work as a whole. It also contains a list of practical suggestions for building further upon the research described in the thesis.

# Contents

# List of Figures

# List of Tables

# Glossary

| Abbreviation | Full |
|---|---|
| AEE | Activity Energy Expenditure |
| BMI | Body Mass Index |
| DIT | Diet-induced Thermogenesis |
| DLW | Doubly-labelled Water |
| DW | Dominant Wrist |
| DXA | Dual-energy X-ray Absorptiometry |
| EE | Energy Expenditure |
| ENMO | Euclidean Norm Minus One |
| HPFVM | High-pass Filtered Vector Magnitude |
| FFQ | Food Frequency Questionnaire |
| LSTM | Long Short Term Memory |
| MET | Metabolic Equivalent Task |
| NDW | Non-dominant Wrist |

| Abbreviation | Full |
|---|---|
| PA | Physical Activity |
| PAEE | Physical Activity Energy Expenditure |
| REE | Resting Energy Expenditure |
| RMR | Resting Metabolic Rate |
| RMSE | Root Mean Squared Error |
| TEE | Total Energy Expenditure |
| VM | Vector Magnitude |
| vSMOW | Vienna-Standard Mean Ocean Water |

Chapter 1

# Introduction

## 1.1 Physical activity

Physical activity has been defined as “any bodily movement produced by skeletal muscles that requires energy expenditure” [18]. In the controlled environment of a laboratory, we have various technologies to measure bodily movement and energy expenditure with high precision, and therefore according to this definition we can achieve a near-perfect measurement of physical activity. However, if we want to study differences in physical activity between individuals and understand what relationship those differences have with health, we need to capture and measure free-living behaviour. Outside of the laboratory, practicality forces us to use a more limited set of tools to observe the physical activity of free-living individuals. The challenge in the latter case is to develop and strengthen the inferences that we make from those limited observations, with the ultimate goal of achieving the precision currently only attainable in the laboratory.

When measuring an individual in free-living conditions, we have to consider several factors. The measurement must be unintrusive and minimise the burden upon the participant, otherwise reactivity issues or lower protocol adherence may compromise the reliability and quality of the dataset. Ease of administration and data collection by the researchers on a large scale is also important. Of course, monetary cost is always an unavoidable concern. Over the last few years, these considerations have lead to a widespread adoption of using accelerometers as an objective way to capture physical activity in many large studies worldwide. More specifically, the general preference has been to attach one accelerometer to a specific anatomical site, rather than a more burdensome array of sensors scattered around the body.

## 1.2 Accelerometry

An accelerometer periodically measures and records the acceleration forces to which it is subjected; the collected data therefore consists of a list of sequential observations throughout the duration of measurement. In practice, it is usually kept attached to the same anatomical site, most commonly worn on the wrist, hip or thigh. It is important to note that an accelerometer is therefore a device for capturing bodily movement, not energy expenditure; by the strictest of definitions, a measurement of energy expenditure requires direct measurement of the metabolic process of respiration, or related physiological phenomena such as heat exchange, oxygen consumption, or heart rate. However, it is hoped that there is a sufficiently strong relationship between bodily movement and energy expenditure that an acceleration trace from a wearable sensor is informative enough that we can accurately estimate energy expenditure from accelerometry signals alone.

The earliest generation of accelerometers were hampered by various hardware limitations, such as insufficient battery and storage capacity to comprehensively record their acceleration measurements. Their solution was to perform live processing of the measured acceleration signal, and to store a summary of the signal at less frequent intervals. The commercial companies responsible for these devices have historically been less than forthcoming about the specifics of their methodologies, which has created uncertainty regarding the mapping between the true bodily movement and the measured value.

The latest generation of accelerometers benefit from modern innovations in electrical hardware, and are capable of recording the raw acceleration forces as measured by the device in all three dimensions - this is often referred to as triaxial acceleration. This has moved the responsibility of all signal processing to a later stage, and has created the

opportunity to revisit and improve upon the methodologies used to process and utilise the recorded data.

For illustrative purposes, Figure 1.1 contains several example triaxial acceleration traces measured at the non-dominant wrist during certain activities, each ten seconds long.


Figure 1.1: Example ten-second long triaxial acceleration traces, measured at the non-dominant wrist during various activities.

### 1.2.1 Processing of triaxial acceleration data

Isolating human movement intensity

A raw triaxial acceleration signal typically contains three main components; gravitational acceleration, acceleration due to human movement, and noise. Without additional information (specifically, gyroscopic information), it is not possible to know exactly how much of each component is contributing to each of the three axes; it is therefore com-

mon practice to collapse the three axes to a unidimensional signal and approximate the removal of gravity from it [82]. The instantaneous acceleration intensity at time $t$ can be characterised by calculating the Vector Magnitude (VM) of the three axes: $VM_t = ({X_t}^2 + {Y_t}^2 + {Z_t}^2)^{0.5}$, where $X_t$ refers to one of the locally-defined directional axes at time $t$, etc.

In a typical acceleration signal collected using a body-worn sensor in free-living conditions, the largest component of this signal is gravity, which by definition is 1 g (gravities) or approximately 9.8 m·s$^{-2}$. Vector Magnitude therefore tends towards approximately 1 g throughout the measurement. It will only be below 1 g if the device is in free-fall or is otherwise accelerated towards the source of gravity (Earth).

There are two favoured ways of removing gravity [82]; one method simply subtracts 1 g from the signal, and truncates any resulting negative values to zero, commonly referred to as Euclidean Norm Minus One (ENMO). Another approach is to apply a high-pass Butterworth filter to the signal at a very low frequency such as 0.2 Hertz, effectively treating gravity as a low-frequency component which can be filtered out, which is called High-Pass Filtered Vector Magnitude (HPFVM). Figure 1.2 shows ENMO and HPFVM inferred from the same triaxial trace measured at the wrist during walking.

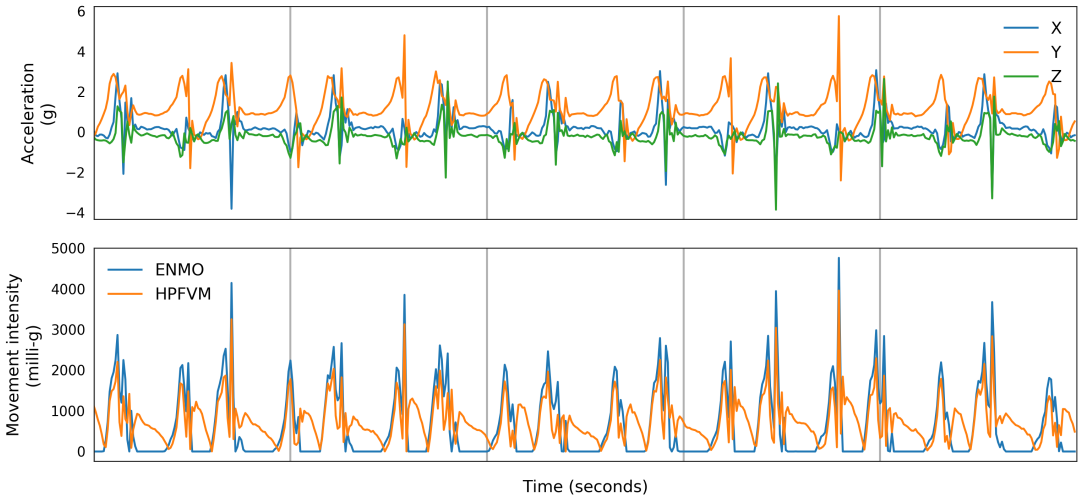


Figure 1.2: Example movement intensity inferences derived from triaxial acceleration data during a repetitive activity over five seconds.

### Accelerometer calibration

An accelerometer contains many electrical and mechanical components, which are subject to manufacturing imperfections or inconsistencies between devices. This can make the device particularly sensitive to acceleration in one axis versus another, or it can introduce a bias wherein the device consistently measures an additional fixed amount of acceleration in a particular direction.

Autocalibration is the process of automatically adjusting the measured acceleration signal in order to compensate for these problems [51, 80]. When the device is stationary, we can reliably assume that the dominant signal is gravity, with a strength of approximately 1 g; therefore we can assume that there is a calibration error if the vector magnitude signal systematically deviates from 1 g when it is still. For all triaxial observation vectors at time $t$ $(X_t,\ Y_t,\ Z_t)$ where $({X_t}^2 + {Y_t}^2 + {Z_t}^2)^{0.5} \neq 1$, we can normalise the vector such that $({\hat{X}_t}^2 + {\hat{Y}_t}^2 + {\hat{Z}_t}^2)^{0.5} = 1$. This new vector $(\hat{X}_t,\ \hat{Y}_t,\ \hat{Z}_t)$ is an "idealised" version of the original observation, but appropriately scaled such that gravity is

the correct “strength”. A transformation can then be calculated by deriving a model that optimally adjusts ($X_t$, $Y_t$, $Z_t$) to match ($\hat{X}_t$, $\hat{Y}_t$, $\hat{Z}_t$) for all $t$, which can be achieved by a linear model for each axis ($\hat{X}_t = \alpha + \beta \times X_t$), etc. Additional factors such as device temperature fluctuations can be factored into such a model [80], as this may affect its measurement properties. The resulting correction is subsequently applied to the entire measured signal, resulting in a measurement that has been autocalibrated to local gravity.

### 1.2.2 Inferences made from accelerometry

What follows is a brief summary of the most common summary statistics that are used to describe and quantify objective physical activity records.

#### Volume of physical activity

The most widely recognised and easily understood description of physical activity is total volume of activity over a given time frame. For a fairer comparison between individual records of different lengths, volume of activity is usually divided by the duration of measurement to yield an average activity intensity. Figure 1.3 shows an example accelerometry trace over a day, and the resultant volume and average activity. Any continuous physical activity intensity signal can be used for this style of analysis; for example, it could be a movement signal such as ENMO or HPFVM derived from triaxial accelerometry, or an activity energy expenditure signal.

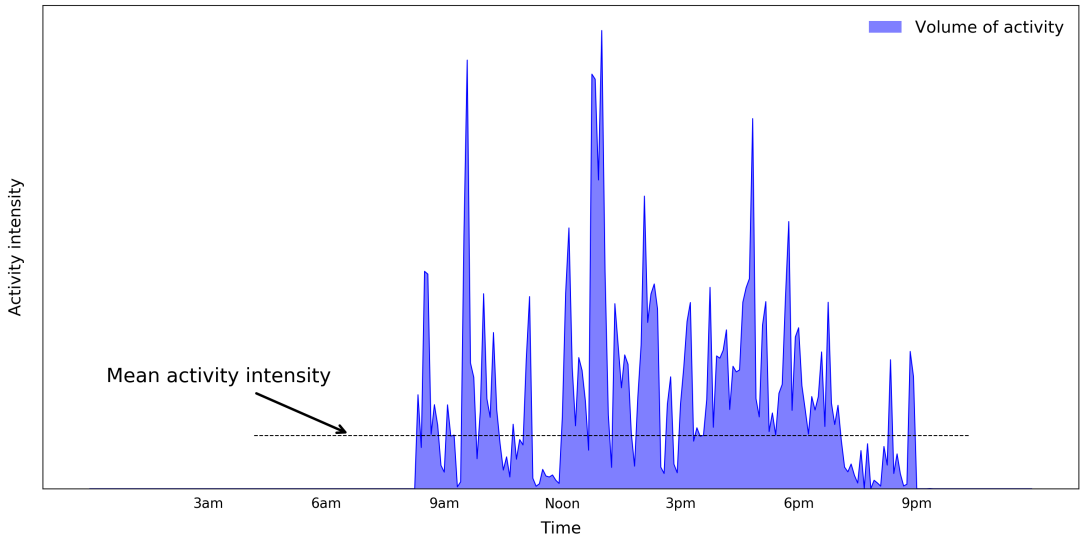


Figure 1.3: An example physical activity trace over one day.

### Intensity distribution

Energy expenditure intensity is sometimes described in METs (Metabolic Equivalent Tasks), where 1 MET is equal to energy expenditure at rest. Traditional labels have emerged for multiples of a MET; anything below 1.5 METs is generally called Sedentary, between 1.5 to 3 METs is Light, 3 to 6 METs is Moderate, and anything above is Vigorous. The intensity distribution of an energy expenditure signal (such as may be inferred from accelerometry) can be described by counting the observations that fall within these MET thresholds. A MET is not a standard unit because resting energy expenditure is variable between individuals [40], but for these purposes it is common to ignore this variation and assume a "standard MET" of 3.5 ml·min$^{-1}$·kg$^{-1}$ of Oxygen consumption. Figure 1.4 shows the example activity trace given in Figure 1.3, summarised according to its classical intensity distribution.

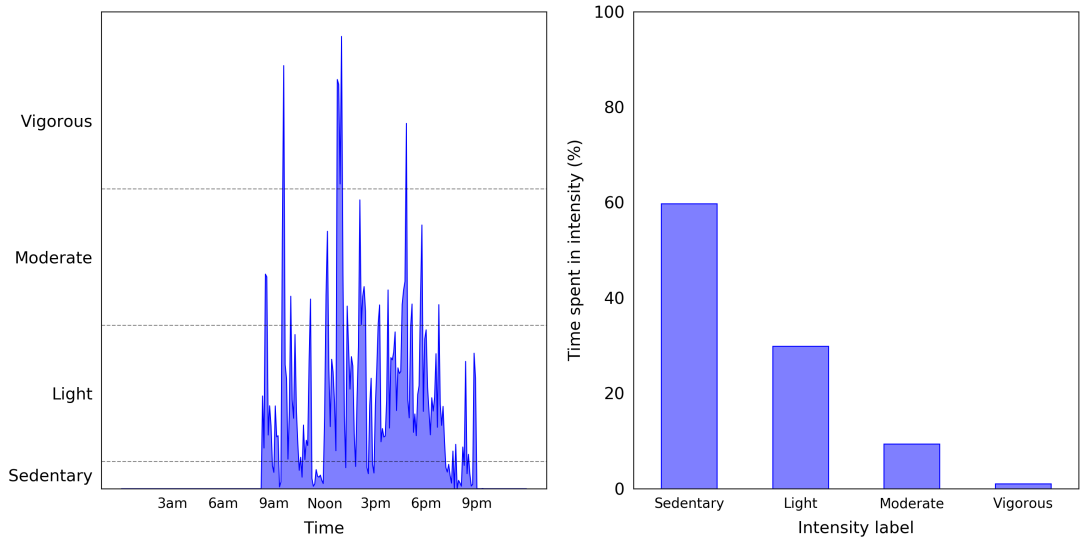


Figure 1.4: A physical activity trace and its corresponding classical intensity distribution.

An alternative approach to characterising the intensity distribution is to place much less emphasis on distinguishing between the traditional categories, and “bin” the intensity in a more systematic way, with a greater resolution [22, 8]. In other words, rather than describing intensity with the four variables representing time spent in Sedentary, Light, Moderate and Vigorous, we could calculate a hundred or even a thousand variables capturing time spent above or below a very specific threshold, as illustrated in Figure 1.5. Such a description is less contingent upon the validity of translating the measurement to energy expenditure, as no special significance is assigned to each individual bin. However, this larger array of variables requires a more careful statistical interpretation, as there is naturally a greater risk of falling victim to multiple-testing problems.

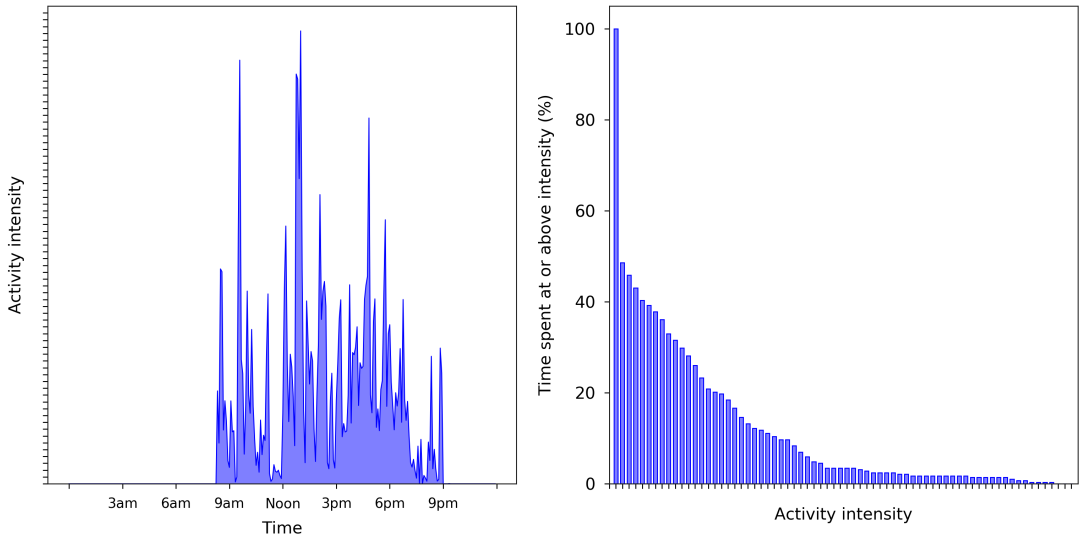


Figure 1.5: A physical activity trace and its corresponding cumulative intensity distribution.

### Bouts

When quantifying an intensity distribution alone, observations of the target intensity are counted unconditionally and equally. This concept can be refined and made more specific by examining *sustained* activity within the specified range, which is referred to as a "bout" of activity. Typically, bouts are described in the form of time spent in a specific intensity range, in bouts of a minimum or maximum duration. For example, a single variable might count the time spent in Moderate intensity or above, in bouts whose duration exceeds 10 minutes.

In practice, current research questions on the topic of bouts generally focus on the two extremes of the intensity spectrum. Some are interested in prolonged bouts of sedentary time, which is suspected to be deleterious [38]. Others are interested in prolonged Moderate and Vigorous intensities, which is suggestive of deliberate exercise and suspected to be beneficial [54, 4], perhaps enough to redeem those at risk by

being sedentary [26]. The physical activity guidelines adopted by many countries and organisations specify that adults should accumulate 150 minutes per week spent in at least Moderate intensity activity in bouts lasting at least 10 minutes [16].

Placing importance on bouts of activity is implicitly to place importance on the specific ordering of activites performed. An intensity distribution alone makes no distinction between an hour long walk and twelve separate five minute walks, which certainly represent different behavioural patterns. There is a complicated relationship between the resolution of the acceleration signal and bouts of activity [63]; a higher resolution signal is more sensitive to momentary outlying intensity values which will terminate a bout, leading to fewer continuous bouts being detected.

### Posture

The pose of the body is a phenomenon of interest to physical activity researchers, as it can potentially be used to provide new context to other measurements of physical activity [68]. Interest in this domain has been encouraged by the consensus statement on the definition of sedentary behaviour, which included the sitting posture as a defining characteristic [17].

Under static conditions, the presence of gravity distributed over the three axes of an acceleration signal allows us to calculate the angular pose in which the device is resting, using trigonometric equations. This observation has been succesfully exploited by commercial devices intended to be attached to specific body parts such as the thigh, which can then be used to quantify behavioural constructs such as sitting time, when attached appropriately [32]. This inference gets progressively less accurate as more human movement is added to the signal; the only way to overcome this is to add a triaxial gyroscope to make the device an inertial measurement unit, but they currently consume too much battery power to be of practical use in the measurement of free-

living individuals. Figure 1.6 illustrates an example pitch and roll signal inferred from a wrist measurement during walking.

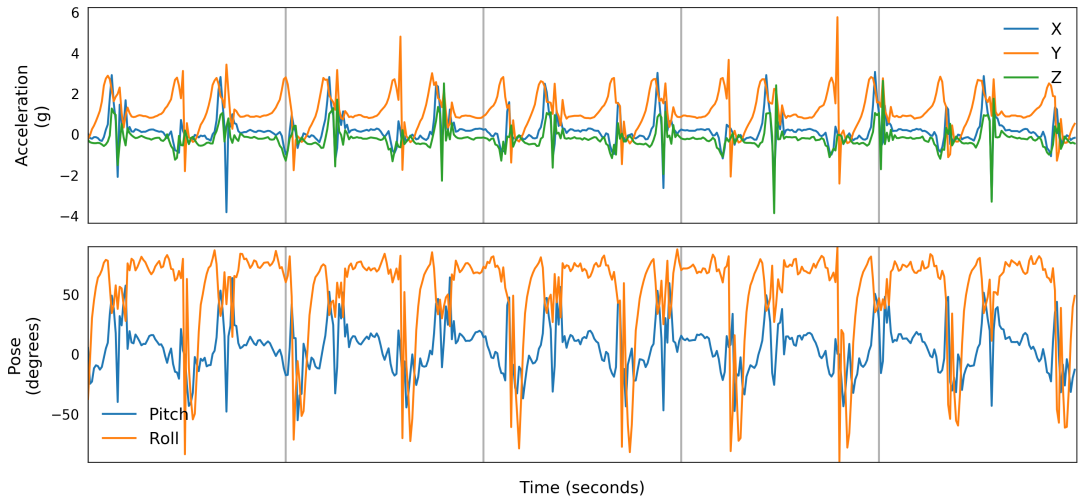


Figure 1.6: Example postural inferences derived from triaxial acceleration data during a repetitive activity over five seconds.

### Activity type

Activity can be characterised based on the descriptive label we would assign it, such as "sitting", "walking" or "running". This has great utility because it is inherently easier to communicate to study participants and the general public. For example, a public health message stating people should "walk for 20 minutes a day" is easier to understand and implement than "expend energy at a rate exceeding 142 $J \cdot min^{-1} \cdot kg^{-1}$ for 20 minutes a day".

The "type" dimension of physical activity has traditionally been captured by instruments such as diaries and logs, where a participant is expected to complete an activity log whilst still going about their daily life, which can be quite onerous and disruptive. The emergence of high resolution triaxial accelerometers has created the opportunity to replace logs and do this automatically at scale; the field of human activity recognition

aims to use the raw data (such as shown in Figure 1.1) to recognise the distinct and reproducible patterns which identify the activity it contains. Human activity recognition has seen a resurgence since the advent of deep learning [49], and the current state-of-the-art utilises deep neural networks to map raw signal features to activity labels [35, 33].

It has been argued that the current most significant challenge to advancing human activity recognition is the relative scarcity of labelled training data [65]. Training classifiers requires an appropriately sized dataset which reflects the diversity and complexity of the problem, and collecting such large scale data (many people, many activities) has thus far proven impractical or infeasible by current methods.

## 1.3 Purpose of this thesis

A growing number of studies have used single body-worn sensors such as accelerometers to record high resolution raw acceleration data in order to record the physical activity of their participants during free-living. In the context of a growing worldwide obesity epidemic and related metabolic disorders, there is a need to understand these physical activity measurements in terms of their associated activity energy expenditure. The aim of this thesis was to build and evaluate an energy-based interpretive framework for these signals, and for those to be collected in the future.

Estimation models are required which predict activity energy expenditure intensity from raw acceleration data, yielding an energy expenditure signal which can subsequently be described in terms of volume, intensity distribution, and bouts. Furthermore, those models need to be evaluated in a manner appropriate to their intended usage, which is the assesssment of activity energy expenditure in a diverse range of free-living individu-

als. Interpreting the results of these models requires a thorough characterisation of the estimation error structure, and an understanding of which factors and circumstances may affect estimation performance.

## Chapter 2

# Estimation of physical activity energy expenditure during free-living from wrist acceleration intensity

The following chapter is adapted from published work:

T White et al., PLoS One, 2016 [91].

*Introduction: Wrist-worn accelerometers are emerging as the most common instrument for measuring physical activity in large-scale epidemiological studies, though little is known about the relationship between wrist acceleration and physical activity energy expenditure (PAEE).*

*Methods: 1695 UK adults wore two devices simultaneously for six days; a combined sensor and a wrist accelerometer. The combined sensor measured heart rate and trunk acceleration, which was combined with a treadmill test to yield a signal of individually-calibrated PAEE. Multi-level regression models were used to characterise the relationship between the two time-series, and their estimations were evaluated in an independent holdout sample. The linear relationship between PAEE and BMI was described separately for each source of PAEE estimate (wrist acceleration models and combined-sensing).*

*Results: Wrist acceleration explained 44–47% between-individual variance in PAEE, with RMSEs between 34–39 J·min$^{-1}$·kg$^{-1}$. Estimations agreed well with PAEE in cross-validation (mean bias [95% limits of agreement]: 0.07 [-70.6; 70.7] J·min$^{-1}$·kg$^{-1}$) but overestimated in women by 3% and underestimated in men by 4%. Estimation error was inversely related to age (-2.3 J·min$^{-1}$·kg$^{-1}$ per decade) and BMI (-0.3 J·min$^{-1}$·kg$^{-1}$ per kg·m$^{-2}$). Associations with BMI were similar for all PAEE estimates (approximately -0.08 kg·m$^{-2}$ per J·min$^{-1}$·kg$^{-1}$).*

*Discussion: A strong relationship exists between wrist acceleration and PAEE in free-living adults, such that irrespective of the objective method of PAEE assessment, a strong inverse association between PAEE and BMI was observed.*

## 2.1 Introduction

Physical activity (PA) is important for the prevention of several chronic diseases such as diabetes, cardiovascular disease, and certain cancers [50]. However, there is uncertainty about the dose-response relationships as well as the prevalence of the exposure, owing to difficulty in assessing habitual physical activity accurately [86]. Several methods now exist but wrist accelerometry is becoming a more common objective measure of habitual physical activity in large-scale epidemiological studies [10, 78], due to its relative low cost and high acceptability by study participants. This necessitates a better understanding of the relationship between wrist acceleration and other measures of physical activity so that estimates of prevalence and disease relationships can be compared between populations assessed using different methods.

A recent consensus statement expressed the imminent need for harmonisation of accelerometry data collected in free-living adults [92]. The current lack of comparability between measurement modalities limits possibilities of assessing the global prevalence of physical activity, or pooling data from multiple sources to better understand its relationship with disease. For example, a meta-analysis aiming to determine whether physical activity attenuates the effect of the FTO gene on obesity risk was forced to dichotomise physical activity (active or inactive) across the multitude of exposure measures [46]; while this was sufficient to confirm the existence of an interaction, it was not possible to determine what dose of activity was necessary to protect against the deleterious FTO variant.

An important component of physical activity is its associated increase in energy expenditure (PAEE), which, if captured during free-living in high time resolution, produces intensity time-series data that can be used to describe a person's behavioural profile. A number of previous studies have validated wrist acceleration derivatives against gold-

standard measures of energy expenditure, such as the doubly-labelled water (DLW) method [83] and indirect calorimetry from respired gas analysis [29]. However, the high cost of DLW has prohibited such work in large population samples, and the nature of the measurement only allows the exploration of total activity volume, rather than the underlying intensity time-series. Breath-by-breath analysis does provide intensity time-series data but the method is not a feasible solution for monitoring energy expenditure in free-living. While laboratory-based comparisons have utility in elucidating the relationships between wrist acceleration and energy expenditure during specific activities, such experiments are unlikely to adequately capture the full spectrum of human activities in representative proportions, and we remain ill-equipped to recognise different activity types in free-living records.

The purpose of this study was to complement existing validation studies by building predictive models of classic PA measures from wrist acceleration derivatives, using both acceleration of the trunk and PAEE collected in free-living as criteria. We then evaluate the derived models by cross-validation in age, sex, and BMI strata, and finally investigate if model performance translates into valid methods of harmonisation by examining their association with obesity, compared to that of the criterion measure.

## 2.2 Methods

This dataset was part of the Fenland Study [61], an ongoing prospective cohort study designed to identify the behavioural, environmental and genetic causes of obesity and type 2 diabetes. Participants were recruited to attend one of three clinical research facilities in the region surrounding Cambridge, UK. All participants provided written informed consent and the study was approved by the local ethics committee (NRES Committee – East of England Cambridge Central) and performed in accordance with the Declaration

of Helsinki.

A subsample of 1,695 participants were asked to wear two devices simultaneously; a combined heart rate and movement sensor (Actiheart, CamNtech, Cambridgeshire, UK), which measured heart rate and uniaxial acceleration of the trunk in 15-second intervals [11], and a wrist accelerometer (GeneActiv, ActivInsights, Cambridgeshire, UK) worn on the non-dominant wrist, which recorded triaxial acceleration at 60 Hertz. Participants were asked to wear the monitors for 6 complete days and advised that both monitors were waterproof and could be worn continuously including during showering and sleeping.

At the clinic visit, prior to the free-living monitoring period, participants performed a ramped treadmill test to establish their individual heart rate response to a submaximal test [13]. These measurements produced calibration parameters to inform a branched equation model of PAEE [12], which has been validated against intensity from indirect calorimetry [76, 77]. Following pre-processing of the heart rate data collected during free-living to eliminate potential noise [75], the branched equation model was applied to calculate instantaneous PAEE (J·min$^{-1}$·kg$^{-1}$). This methodology has been successfully validated against PAEE from DLW in several populations [6, 84], including a sample of UK men and women where it was shown to explain 41% of the variance in free-living PAEE and with no mean bias [14].

The raw triaxial wrist acceleration data was auto-calibrated to local gravitational acceleration (in g) using a method described elsewhere [80]. The calibrated acceleration was then used to calculate Vector Magnitude (VM) per sample: $VM(X,Y,Z) = \sqrt{(X^2+Y^2+Z^2)}$. VM, or Euclidean Norm, can be interpreted as the magnitude of acceleration the device was subjected to at each measurement, including gravitational acceleration. There will also be a potential noise component in the high frequency domain, which we filtered out by a 20 Hertz low-pass filter. In the present study, we cal-

culated two derivatives of VM, both aiming to remove the gravity component from the signal in order to isolate the activity-related acceleration component; 1) Euclidean Norm Minus One (ENMO) subtracts 1 g from VM and truncates the result to zero at sample level, whereas 2) High-Pass Filtered Vector Magnitude (HPFVM) applies a high-pass filter to the VM signal at 0.2 Hertz, therefore treating gravity as a low-frequency component to be filtered out. These two signals, ENMO and HPFVM, are both plausible approximations of acceleration as a result of human movement, and are the primary descriptions of wrist acceleration used in the following analyses.

Non-wear detection procedures were applied to both the wrist acceleration [83] and combined-sensing traces [14], and any such non-wear periods were excluded from these analyses. Briefly, non-wear in the wrist acceleration data was defined as time periods where the standard deviation of acceleration in each axis fell below 13mg for longer than 1 hour, and non-wear in the combined sensing data was defined as extended periods of non-physiological heart rate concomitantly with extended ($\geq$ 90min) periods of zero movement registered by the accelerometer.

All signals were down-sampled from their original resolutions (60 Hertz for wrist acceleration intensity, and 0.067 Hertz for individually-calibrated energy expenditure) to a common time resolution of one observation per 5 minutes, by taking the mean average of all observations within each 5 minute window, an example of which is shown in Figure 2.1. This was chosen as an appropriate window length based on a variety of competing considerations. Firstly, the time-lagged physiological response of heart rate to movement precludes an instantaneous comparison and necessitates a physiologically appropriate time buffer. Secondly, due to hardware limitations and initialisation conditions, we could not guarantee a perfect time synchronisation between the two monitors. Finally, maintaining the highest possible time resolution within these constraints preserves the most variance in the intensity time-series, and maximises the number of observations

in the dataset. The models derived in this work (described in detail below) exclusively use time-invariant signal features such as arithmetic means; this means that they are robust to changes in window size, and it is therefore equally appropriate to use them to estimate hourly or daily outputs from hourly or daily inputs.

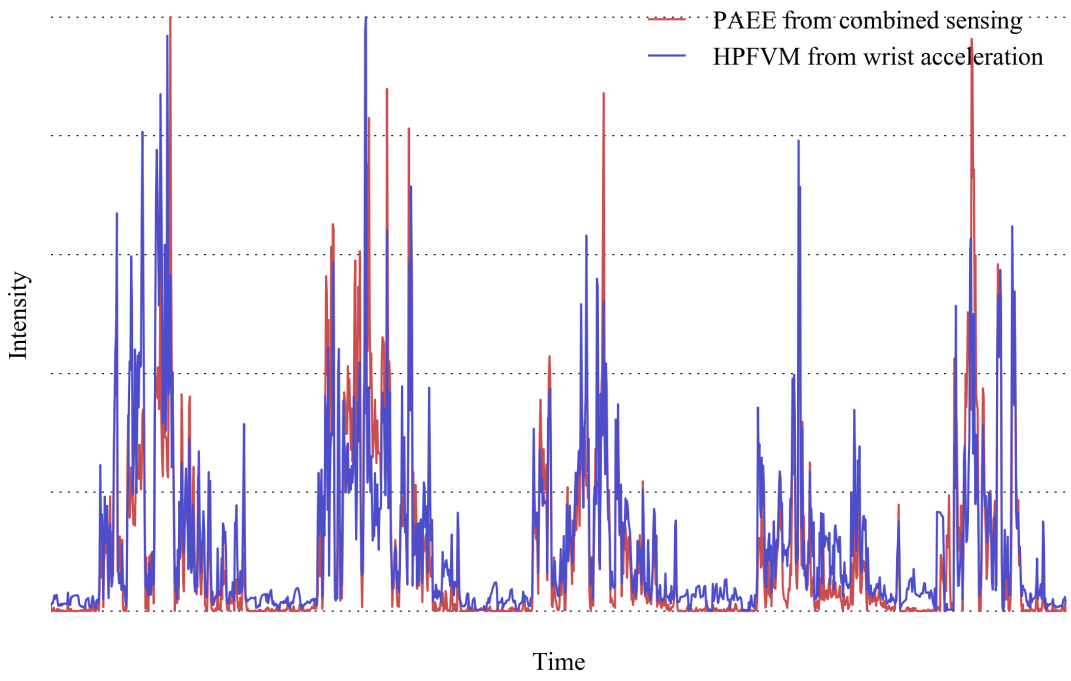


Figure 2.1: Example of a simultaneous PAEE and wrist acceleration signal over 5 days.

Multi-level linear regression models were designed to independently predict PAEE and trunk acceleration from wrist acceleration. Four models were tested: a linear and quadratic model for each of ENMO and HPFVM.

$$\alpha + \beta_1(ENMO)$$

$$\alpha + \beta_1(ENMO) + \beta_2(\sqrt{ENMO}) + \beta_3(ENMO^2)$$

$$\alpha + \beta_1(HPFVM)$$

$$\alpha + \beta_1(HPFVM) + \beta_2(\sqrt{HPFVM}) + \beta_3(HPFVM^2)$$

The models were derived in a randomly chosen subset containing 60% of people in the cohort (n=1,050) and evaluated in the remaining 40% (n=645). Model performance was assessed using within- and between-individual explained variances (Pearson's coefficient) and Root Mean Squared Error (RMSE) metrics, as determined from ANOVA repeated measures modelling specified with random effects at the participant level. After assessing the performance of these models on the test dataset as a whole, we selected the strongest model and tested for differential bias by sex, age and BMI categories of under/normal-weight, overweight, and obese ($<$25, $>$25 and $<$30, $>$30 kg/m$^2$, respectively) within the test dataset. All statistical tests were performed in Stata version 14 (StataCorp, Texas, USA).

In order to test the epidemiological utility of the derived models, we examined the associations between BMI and PAEE in the test dataset (n=645). Using our criterion PAEE measure, we first characterised the linear dose-response relationship with BMI, adjusting for age and sex. We then repeated this analysis using predicted PAEE from each of the derived prediction models, and compared the beta coefficients and 95% confidence intervals to those using criterion PAEE.

## 2.3 Results

A description of the population sample included in this analysis is given in Table 2.1. In total, 1,752,287 valid 5-minute windows from 1,050 individuals were included in the training dataset; the median number of observations per individual was 1,738, equating to just over 6 days. Mean PAEE across the sample was 36.4 J·min$^{-1}$·kg$^{-1}$, with higher average in men than women (38.1 and 34.7, respectively). Mean wrist accel-

eration according to both the ENMO and the HPFVM metrics was similar in men and women.

| | Train | | Test | |
|---|---|---|---|---|
| | Men | Women | Men | Women |
| N | 499 | 551 | 305 | 340 |
| Age (years) | 49.68 (7.29) | 50.07 (7.00) | 51.58 (6.97) | 49.90 (7.29) |
| Height (m) | 1.78 (0.06) | 1.63 (0.06) | 1.77 (0.06) | 1.64 (0.06) |
| Weight (kg) | 85.85 (14.02) | 69.97 (13.05) | 85.50 (13.01) | 69.44 (12.85) |
| BMI ($kg \cdot m^{-2}$) | 26.98 (4.08) | 26.05 (4.86) | 26.99 (3.90) | 25.68 (4.71) |
| PAEE ($J \cdot min^{-1} \cdot kg^{-1}$) | 39.10 (16.44) | 34.51 (13.50) | 37.41 (15.57) | 35.54 (14.46) |
| Trunk ACC ($m \cdot s^{-2}$) | 0.12 (0.05) | 0.13 (0.06) | 0.12 (0.05) | 0.13 (0.05) |
| Wrist ENMO (milli-g) | 32.15 (9.28) | 31.25 (8.12) | 31.36 (9.17) | 31.59 (7.92) |
| Wrist HPFVM (milli-g) | 49.17 (11.73) | 47.75 (10.35) | 47.89 (11.81) | 48.09 (10.23) |

Table 2.1: Summary statistics of the cohort, provided separately for the training and test datasets, by sex. Values given are mean (standard deviation).

The derived regression models for PAEE and trunk acceleration are given in Tables 2.2 and 2.3, respectively, and their overall estimation performance is shown in Figure 2.2. Figures 2.3 and 2.4 visualise the relative density of points between wrist acceleration intensity and PAEE using hexagonal heatmaps (due to the large number of observations, traditional scatterplots are infeasible). Between-individual explained variance in trunk acceleration was between 51% and 56% for all models. For PAEE, there were only minor differences between models in terms of explained variance, ranging from 44% to 47%; but there were slightly more pronounced differences in RMSE, ranging from to 38.8 $J \cdot min^{-1} \cdot kg^{-1}$ at worst to 34.4 at best. (For reference, 1 standard MET is 71 $J \cdot min^{-1} \cdot kg^{-1}$.) Model 4 was the strongest model to discriminate activity intensity levels, as it yielded the lowest RMSE for both criterion measures.

| Model | Formula to predict PAEE ($J \cdot min^{-1} \cdot kg^{-1}$) | Within-$r^2$ | Between-$r^2$ | RMSE ($J \cdot min^{-1} \cdot kg^{-1}$) |
|---|---|---|---|---|
| 1 | $5.01 + 1.000(ENMO)$ | 0.60 | 0.44 | 38.8 |
| 2 | $-10.58 + 1.1176(ENMO) + 2.9418(\sqrt{ENMO}) + 0.00059277(ENMO^2)$ | 0.66 | 0.44 | 35.7 |
| 3 | $-4.65 + 0.8537(HPFVM)$ | 0.68 | 0.47 | 35.0 |
| 4 | $-1.25 + 1.1353(HPFVM) - 2.4281(\sqrt{HPFVM}) - 0.00040270(HPFVM^2)$ | 0.69 | 0.47 | 34.4 |

Table 2.2: Derived regression models of physical activity energy expenditure from non-dominant wrist movement intensity.

| Model | Formula to predict trunk acceleration ($m \cdot s^{-2}$) | Within-$r^2$ | Between-$r^2$ | RMSE ($m \cdot s^{-2}$) |
|---|---|---|---|---|
| 1 | $-0.057 + 0.0060321(ENMO)$ | 0.59 | 0.51 | 0.245 |
| 2 | $0.0423 + 0.0087(ENMO) - 0.03860(\sqrt{ENMO}) - 0.00000129(ENMO^2)$ | 0.59 | 0.52 | 0.243 |
| 3 | $-0.097 + 0.0047835(HPFVM)$ | 0.57 | 0.53 | 0.251 |
| 4 | $0.114 + 0.007367(HPFVM) - 0.057613(\sqrt{HPFVM}) - 0.000001428(HPFVM^2)$ | 0.62 | 0.56 | 0.234 |

Table 2.3: Derived regression models of trunk acceleration.

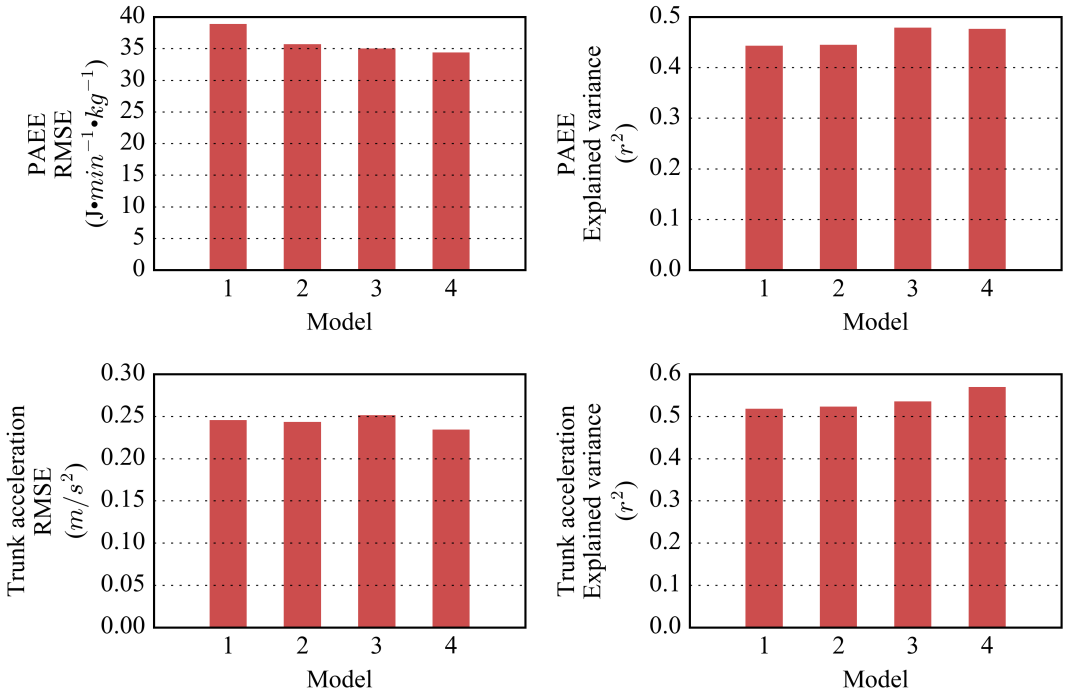


Figure 2.2: Performance of the four models of wrist acceleration. The explained variance shown is between-individual explained variance from ANOVA repeated measures.

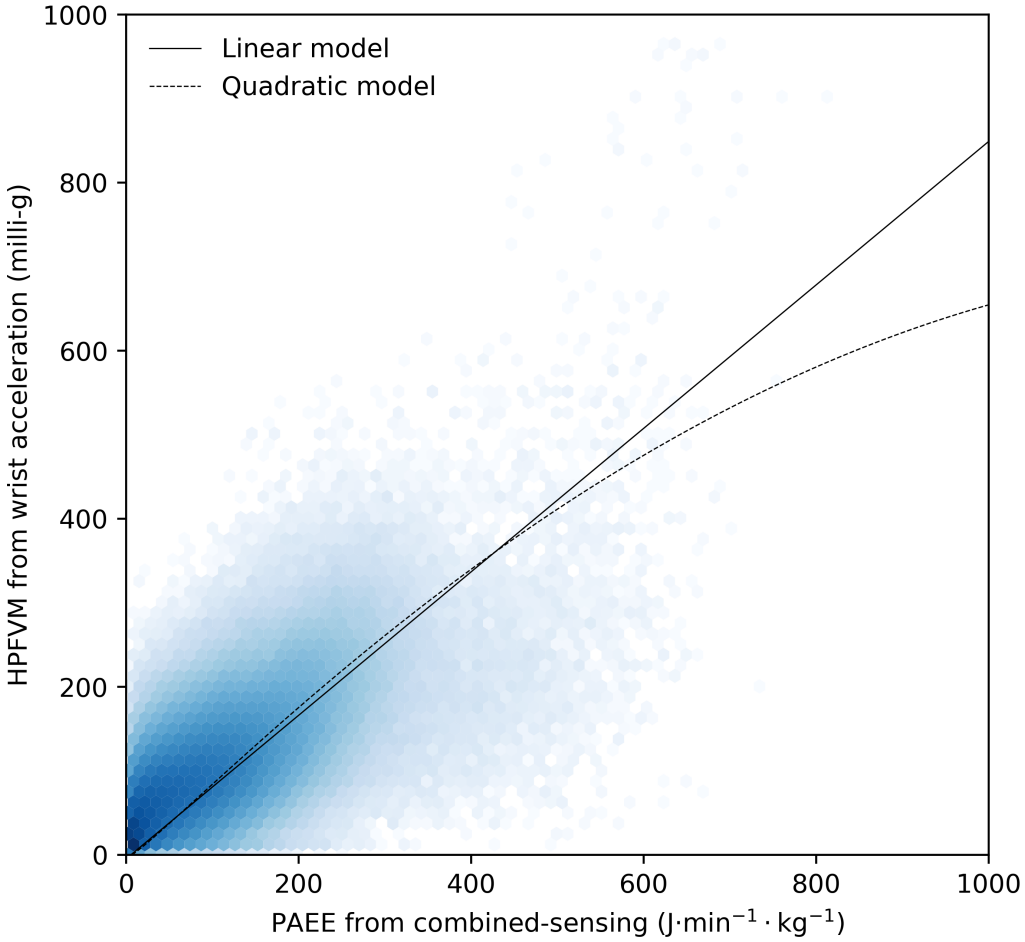


Figure 2.3: Hexagonal heatmaps showing the density of wrist acceleration intensity (HPFVM) relative to PAEE, with the corresponding regression models super-imposed.

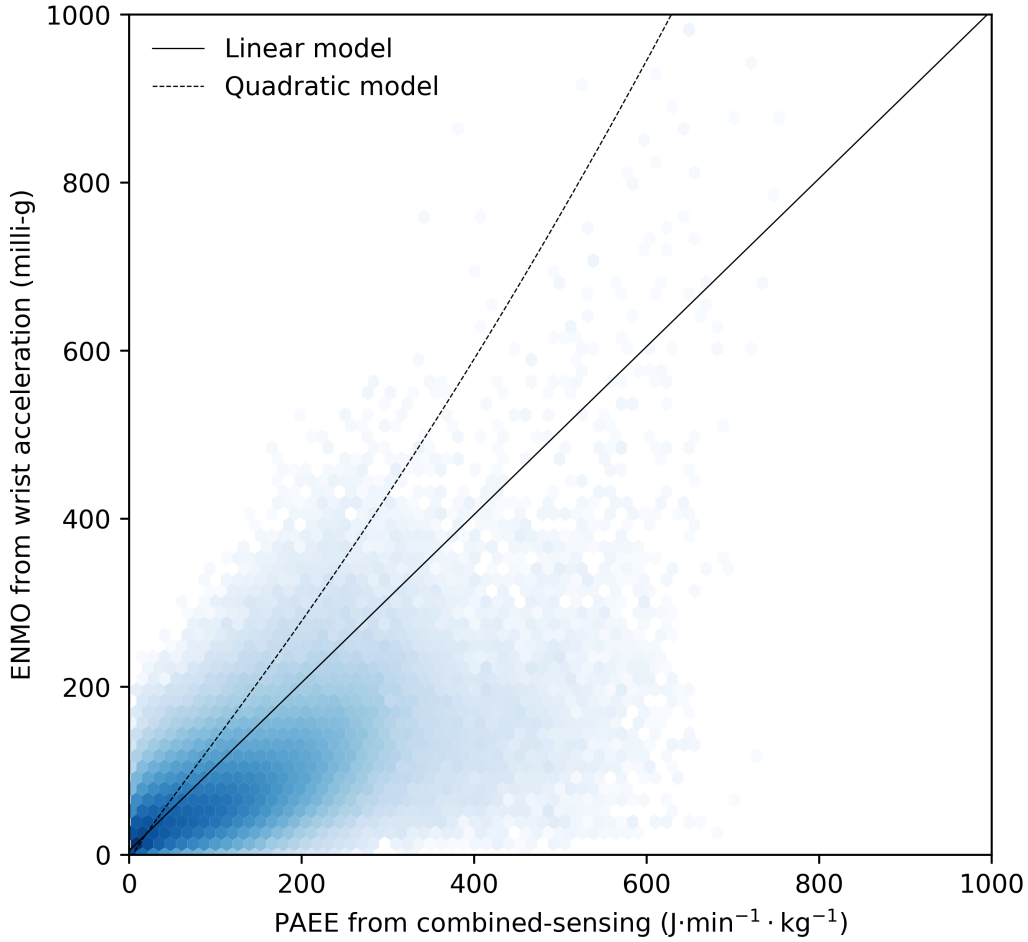


Figure 2.4: Hexagonal heatmaps showing the density of wrist acceleration intensity (ENMO) relative to PAEE, with the corresponding regression models super-imposed.

Model 1 contains a linear term for ENMO, which is the most common signal derivative in current use for wrist acceleration data; it explained 44% of the between-individual variance in PAEE and has a RMSE of just above 0.5 METs. The family of models using HPFVM as the wrist acceleration metric generally outperformed their ENMO counterparts by 2 to 3% in predicting both trunk acceleration and PAEE. The quadratic mod-

els outperformed their linear counterparts, decreasing RMSE by 2 to 8% implying that the relationships between wrist acceleration and both trunk acceleration and PAEE are curvilinear, rather than linear.

Comparing the predictions of model 4 to PAEE from combined sensing in the cross-validation sample (n=645) showed a negligible mean bias (0.07 J·min$^{-1}$·kg$^{-1}$) with 95% limits of agreement between -70.6 and 70.7 J·min$^{-1}$·kg$^{-1}$ (Figure 2.5, panel 1). Stratified by sex, results indicated a 1.2 J·min$^{-1}$·kg$^{-1}$ overestimation in women, and a 1.8 J·min$^{-1}$·kg$^{-1}$ underestimation in men. Age and BMI were centred on their means for this analysis, therefore their coefficients imply a trend from underestimation in the younger and less obese towards overestimation in the older and more obese (0.2 J·min$^{-1}$·kg$^{-1}$ per year relative to mean age, and 0.3 J·min$^{-1}$·kg$^{-1}$ per kg·m$^{-2}$ relative to mean BMI). The distribution of this estimation error is visualised in Figure 2.5 using violin plots and overlaying traditional boxplots; the first panel shows the error distribution in the whole test set, and the remaining panels show error distributions within specific groups within the test set for comparison. It can be seen that estimation error was densely concentrated around zero for all groups, that there were no unusual estimation artefacts, and there were no outstanding differences between any of the groups.

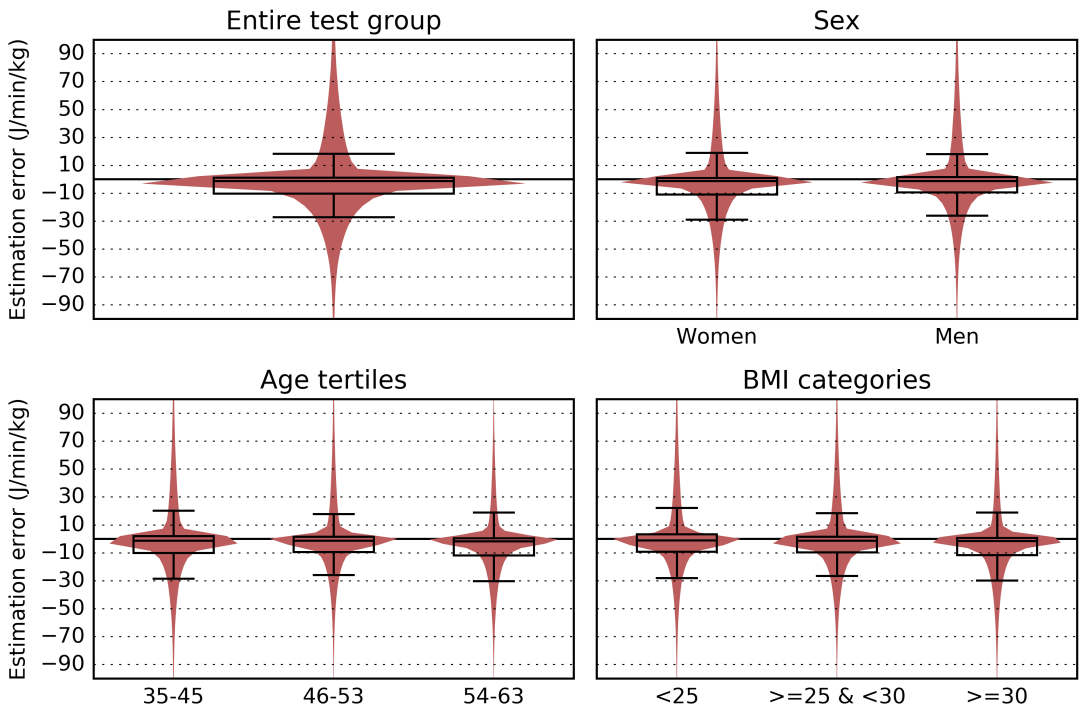


Figure 2.5: Violin plots and boxplots showing the estimation bias of model 4 across the whole test group (top left), by sex (top right), by age tertiles (bottom left) and BMI categories (bottom right).

The association between PAEE and BMI was inverse across all models; the beta coefficients and their respective confidence intervals are visualised in the forest plot in Figure 2.6. All but one of the point estimates from the prediction models fell within the 95% confidence interval of the combined-sensing beta coefficient, and all confidence intervals from the wrist models overlapped the point estimate from combined sensing. The one outlying point estimate was from model 1, the weakest performing model according to other evaluations; however its quadratic counterpart yielded the closest matching beta coefficient of all models.

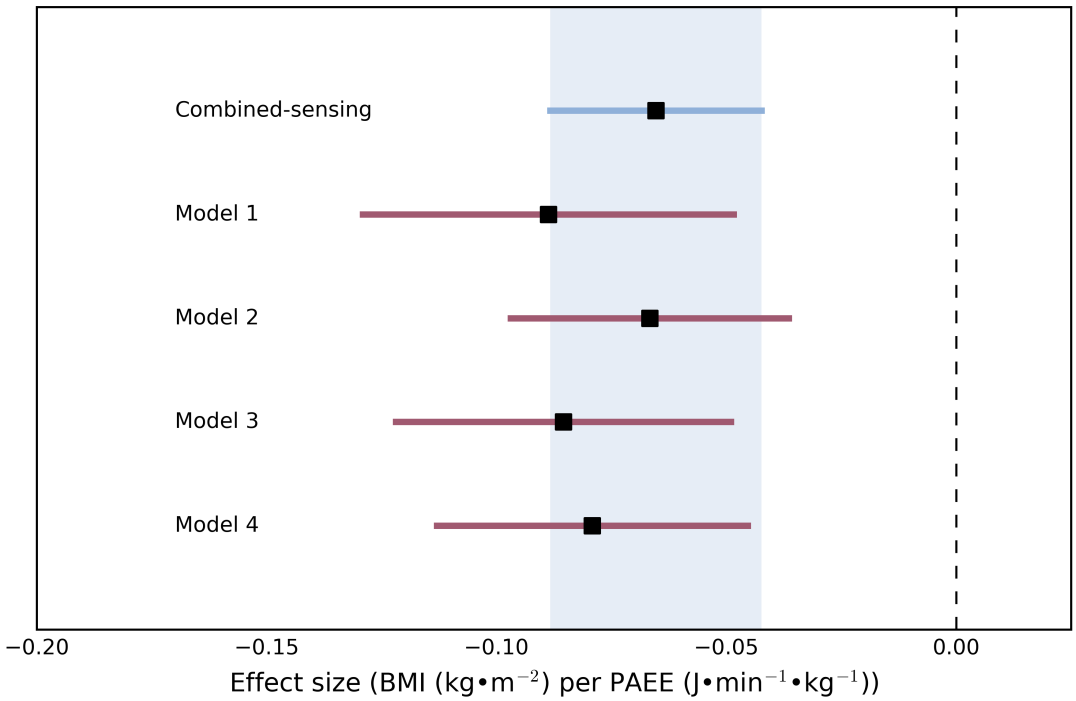


Figure 2.6: Forest plot showing the beta point estimates and their respective confidence intervals of the effect size of PAEE on BMI.

## 2.4 Discussion

To our knowledge, this is the first study to describe the validity of predicting high-resolution free-living PAEE from wrist acceleration in a large population sample of adult men and women, allowing evaluation of model performance in population sub-groups.

Simple models of wrist acceleration intensity were found to explain a high proportion of variance in both PAEE and trunk acceleration, with no evidence of significant difference in bias by age or BMI categories but small opposite biases in men and women (underestimation and overestimation of PAEE, respectively).

The derived equations with non-linear terms were not monotonic; however the non-linear terms responsible for these were statistically significant in all cases. The global maxima of these equations are likely reflective of the highest observed activity levels within the measured population; the data is naturally densely concentrated in the low end of physical activity intensity, and very sparse at the high end, therefore the downward trend at the high end can be considered an artefact of overfitting to the lower end. In practice, an implementation of these equations should truncate the estimates to the global maxima and minima (or zero) where appropriate.

The slightly better performance of HPFVM models compared to ENMO models suggests that applying a high-pass filter to the VM signal may be a more effective approach to the removal of gravity from an acceleration trace. However, the result of this filtering is likely to be dependent upon various signal properties, such as the machine noise level and sampling frequency, and the rotational frequency of human movement with respect to gravity [82].

A traditional validation study would only be able to report the estimation error structure, leaving readers to speculate whether similar associations between a measured behaviour and an outcome would be observed, irrespective of method. We compared the associations between PAEE and BMI as an example; the beta coefficients in the models for predicted PAEE were strikingly close to the beta-coefficient for PAEE from combined sensing, with a strong overlap of the 95% confidence intervals. This final analysis demonstrates that a similar direction and magnitude of relationship between PAEE and BMI can be observed in this population, regardless of whether PAEE is estimated by wrist acceleration or combined-sensing.

The models explored in this analysis only utilised the magnitude of wrist acceleration for prediction, and still achieved strong results. There is potentially a greater explanatory power to be found in the multitude of signal features that are derivable from three-

dimensional acceleration in waveform resolution. Nonetheless, we should be cautiously optimistic that even the basic and robust properties of this easily obtainable and commonly used measure are strongly related to the criterion measure of PAEE from combined sensing.

The validity of these analyses is naturally contingent upon the validity of the criterion measure, individually-calibrated combined sensing of heart rate and trunk acceleration. While it is not considered a gold-standard measurement of PAEE, this estimation method does have established validity of both intensity [76, 77] and PAEE during free-living in the population used for the present evaluation [14], and to our knowledge this study currently represents the largest aggregation of simultaneous wrist acceleration and energy expenditure signals in free-living.

An additional potential limitation of this study is that it is neither nationally or globally representative, but confined to a relatively affluent and culturally homogenous population living in the East of England. The prevalence of many activities, during which wrist acceleration may be more or less representative of PAEE, is likely determined by several factors such as culture, climate, and local landscape, and it is therefore possible that the specific relationships and error structures that we report here may not be universal. Still, our analytical sample comprises both men and women across a wide BMI and activity level range, thus providing a comparative framework for interpreting wrist acceleration data from population studies.

In conclusion, we have demonstrated that a strong relationship exists between PAEE and wrist acceleration. The best performing model explained 47% of the between-individual variance in PAEE with a RMSE of 34 $J \cdot min^{-1} \cdot kg^{-1}$ (0.48 METs) and all prediction models produced similar associations with BMI. Further work should aim to improve upon the accuracy of PAEE prediction using a wider range of the signal feature space, and to explore generalizability in other populations.

## 2.5 Acknowledgements

We are grateful to the participants for giving up their time for this study. We thank all members of the MRC Epidemiology Unit functional groups, including field epidemiology, IT and data management for their contribution to the study, as well as the Fenland principal investigators. Special thanks to Stefanie Hollidge and Lewis Griffiths for their assistance in processing the physical activity data.

## Chapter 3

# Estimating energy expenditure from wrist and thigh accelerometry in free-living adults: a doubly labelled water study

The following chapter is in submission with the International Journal of Obesity.

*Introduction: Many large studies have implemented wrist or thigh accelerometry to capture physical activity, but the accuracy of these measurements to infer Activity Energy Expenditure (AEE) and consequently Total Energy Expenditure (TEE) has not been demonstrated. The purpose of this study was to assess the validity of acceleration intensity at wrist and thigh sites as estimates of AEE and TEE under free-living conditions using a gold-standard criterion.*

*Methods: Measurements for 193 UK adults (105 men, 88 women, aged 40-66 years, BMI 20.4-36.6 kg·m$^{-2}$) were collected with triaxial accelerometers worn on the dominant wrist, non-dominant wrist and thigh in free-living conditions for 9-14 days. In a subsample (50 men, 50 women) TEE was simultaneously assessed with doubly labelled water (DLW). AEE was estimated from non-dominant wrist using an established estimation model, and novel models were derived for dominant wrist and thigh in the non-DLW subsample. Agreement with both AEE and TEE from DLW was evaluated by mean bias, Root Mean Squared Error (RMSE) and Pearson correlation.*

*Results: Mean TEE and AEE derived from DLW was 11.6 (2.3) MJ·day$^{-1}$ and 49.8 (16.3) kJ·day$^{-1}$·kg$^{-1}$. Dominant and non-dominant wrist acceleration were highly correlated in free-living (r=0.93), but less so with thigh (r=0.73 and 0.66, respectively). Estimates of AEE were 48.6 (11.8) kJ·day$^{-1}$·kg$^{-1}$ from dominant wrist, 48.6 (12.3) from non-dominant wrist, and 46.0 (10.1) from thigh; these agreed strongly with AEE (RMSE =12.2 kJ·day$^{-1}$·kg$^{-1}$, r=0.71) with small mean biases at the population level (6%). Only the thigh estimate bias was statistically significantly different from the criterion. When combining these AEE estimates with estimated REE, agreement was stronger with the criterion (RMSE=1.0 MJ·day$^{-1}$, r=0.90).*

*Discussion: In UK adults, acceleration measured at either wrist or thigh can be used to estimate population levels of AEE and TEE in free-living conditions with high precision.*

## 3.1 Introduction

Characterising the energy balance of individuals in free-living conditions requires an accurate assessment of total energy expenditure. Total energy expenditure can be measured with high precision using the doubly labelled water technique [72] but this is an expensive undertaking that requires elaborate sample collection and analysis infrastructure, making it less feasible for large-scale deployment or application in clinical settings. In most people, the largest component of total energy expenditure is resting energy expenditure, which can be predicted from anthropometric information with reasonable accuracy [40, 69]. Diet-induced thermogenesis is less variable and ordinarily constitutes approximately 10% of total energy expenditure [88]. The predominant source of uncertainty in total energy expenditure estimates is the highly variable activity energy expenditure component, which has proven difficult to capture by subjective instruments such as questionnaires [53, 31]. Body-worn sensors such as accelerometers have the potential to provide a relatively cheap and reliable solution to this problem [64], if valid inference models can be devised to estimate activity energy expenditure from the measurements they record.

In recent years, wrist-worn accelerometers have become a popular measurement modality for objectively capturing free-living physical activity in large-scale studies [25, 22, 78]. Devices worn on the wrist are generally considered to be less burdensome for participants than those worn on other anatomical sites [83]. This has led to improved wear protocol adherence and thus to measurements with potentially greater representation of habitual physical activity levels. However, despite their recent increase in popularity, their utility in the estimation of activity energy expenditure has yet to be tested against gold-standard techniques in a sufficiently large sample of free-living men and women [64]. Furthermore, some large studies [22, 78, 25] have committed to measuring only

one of either the dominant wrist or non-dominant wrist, and the relationship between these two measurements also remains understudied.

In previous work, we derived parametric models to estimate activity energy expenditure intensity from non-dominant wrist acceleration (reproduced in Table 2) using a large dataset (n=1050) of simultaneous non-dominant wrist and individually-calibrated combined heart rate and movement sensing signals collected under free-living conditions [91]. We evaluated the models in a large holdout sample (n=645) and found that they explained 44-47% of the variance in activity energy expenditure with no significant mean bias at the population level. However, as this comparison was against a silver-standard measurement of activity volume, these estimation models could be more conclusively validated by integrating the estimated activity energy expenditure signal over time, and assessing agreement of activity volume with a gold-standard criterion such as doubly labelled water. This approach has been used to validate combined heart rate and movement sensing [14, 84, 6], against which the models were originally derived.

Thigh-worn devices have also been used, but have typically been employed in smaller studies to measure time spent in a sitting posture, in order to infer sedentary time. This is possible because the distribution of gravity over the three axes can be interpreted using trigonometry to calculate thigh inclination. However, thigh acceleration has received comparatively little attention as a measure of physical activity intensity during free-living, though it features prominently in activity classification experiments [7]. In epidemiological settings, thigh-worn sensors have been complemented by other sensors with the intention to capture physical activity separately [39].

The primary aim of this study was to describe the absolute validity of a previously established activity energy expenditure prediction model [91] when applied to both wrists, and to evaluate the validity of this estimation in predicting total energy expenditure when combined with a simple anthropometric prediction of resting energy expenditure [40].

The second aim was to use the same approach to derive and validate similar energy expenditure estimation models using thigh acceleration. The third aim was to explore the relationship between the dominant wrist, non-dominant wrist and thigh acceleration measures in free-living, and to derive intensity models to facilitate harmonisation.

## 3.2 Methods

Participants were recruited from the Fenland study, an ongoing cohort described in detail elsewhere [61]. We aimed to recruit participants who had previously indicated that they were interested in participating in future studies, were aged between 40 and 70 years, with a BMI between 20 and 50 kg·m$^{-2}$. Recruitment aimed to balance age, sex and BMI distributions. A total of 193 individuals agreed to participate. Participants were invited to attend an assessment centre on two separate occasions, separated by a free-living period of 9 to 14 days. Ethical approval for the study was obtained from Cambridge University Human Biology Research Ethics Committee (Ref: HBREC/2015.16). All participants provided written informed consent.

Weight was measured to the nearest 0.1 kg using calibrated digital scales (TANITA model BC-418 MA; Tanita, Tokyo, Japan) at both visits. Height was measured to the nearest 0.1 cm using a stadiometer (SECA 240; Seca, Birmingham, UK) at the first clinic visit. Body composition was also measured by DXA (Lunar Prodigy Advanced, GE Healthcare, USA) as part of the Fenland study.

Total energy expenditure was measured by doubly labelled water in 100 of the participants. Prior to the first clinic visit, participants self-reported their current weight, which was used to provide a body-weight specific dose of $^{2}H_{2}^{18}O$ (70 mg $^{2}H_{2}O$ and 174 mg $H_{2}^{18}O$ per kg body weight). Participants brought a baseline urine sample to

their first clinic visit, and a second baseline sample was taken at the clinic visit, prior to dosing. Participants were provided labelled sampling bottles and asked to collect one urine sample per day for the next 9-10 days, at a similar time each day but not the first void of the day. Participants were asked to record the date and time of each measurement on the sample bottle label and separately on a provided timesheet. Participants were asked to store the samples in a container in a cool, dry place, such as a refrigerator, and to return those samples at their second clinic visit at the end of their free-living measurement period. Isotope ratio mass spectrometry ($^{2}$H, Isoprime, GV Instruments, Wythenshaw, Manchester, UK and $^{18}$O, AP2003, Analytical Precision Ltd, Northwich, Cheshire, UK) was used to measure the isotopic enrichment of the samples. All samples were measured alongside laboratory reference standards, previously calibrated against the international standards Vienna-Standard Mean Ocean Water (vSMOW) and Vienna-Standard Light Antarctic Precipitate (vSLAP) (International Atomic Energy Agency, Vienna, Austria). Sample enrichments were corrected for interference according to Craig [21] and expressed relative to vSMOW. Rate constants and pool sizes were calculated from the slopes and intercepts of the log-transformed data, with total $CO_2$ production ($RCO_2$) calculated using the multi-point method of Schoeller [71]. $RCO_2$ was converted to total energy expenditure [27] where the respiratory quotient was informed by the macronutrient composition of the diet (see below).

Resting metabolic rate was measured at the start of both clinic visits during a fifteen-minute rest test by respired gas analysis (OxyconPro, Jaeger, Germany). A seven-breath running median was calculated and the lowest observed average rate over a five minute consecutive window was found, which was scaled down by 6% to compensate for within-day elevation of resting metabolic rates [36]. Basal metabolic rate was also estimated via three different equations which differ in the specific body composition information utilised [40, 60, 87]. Resting energy expenditure was primarily characterised as the nearest measured value to the mean average estimated value, and a further

analysis was conducted using the mean average of the measured values. The final 24-hour resting energy expenditure estimates also included an adjustment for a 5% lower metabolic rate during sleep [3], which was applied according to their reported mean sleep duration.

At the second clinic visit, participants were asked to complete a Food Frequency Questionnaire [9], which was used to estimate dietary intake over the past year. The food frequency data was processed using FETA [59], and the resulting calorie-weighted macronutrient profile was used to calculate the Food Quotient and diet-induced thermogenesis [44]. Diet-induced thermogenesis was normalised by the total energy expenditure to total energy intake ratio, as done previously [14].

At the first clinic visit, participants were fitted with three waterproof triaxial accelerometers (AX3, Axivity, Newcastle upon Tyne, UK); one device was attached to each wrist with a standard wristband, and one was attached to the anterior midline of the right thigh using a medical-grade adhesive dressing. The devices were setup to record raw, triaxial acceleration at 100 Hz with a dynamic range of 8 g (where g refers to the local gravitational force, roughly equal to 9.81 $\mathrm{m{\cdot}s^{-2}}$). Participants were asked to wear them continuously for the following 8 days and nights whilst continuing with their usual activities. They were also asked to record their main sleep using a sleep diary throughout the free-living period.

The signals were resampled from their original irregularly timestamped intervals to a uniform 100 Hertz signal by linear interpolation, and then calibrated to local gravity using a well-established technique [80], [51], without adjustment for temperature changes within the record. Periods of nonwear were identified as windows of an hour or more wherein the device was inferred to be completely stationary [83], where stationary is defined as standard deviation in each axis not exceeding the approximate baseline noise of the device itself (10 milli-g). Vector Magnitude (VM) was then calculated from

the three axes (VM (X,Y,Z) = $(X^2 + Y^2 + Z^2)^{0.5}$), from which two acceleration intensity metrics were derived [82]; Euclidean Norm Minus One (ENMO) subtracts 1 g from VM and truncates any negative results to 0, and High-Pass Filtered Vector Magnitude (HPFVM) applies a fourth-order high-pass filter to the signal at a 0.2 Hertz cut-off (3 dB). These analyses were performed using pampro v0.4.0 [90].

In the non-doubly labelled water group (n=93), multi-level linear regression with random effects at the participant level was used to characterise each of the pairwise relationships between dominant wrist, non-dominant wrist and thigh acceleration intensity using synchronised 5-minute level data from each source. We used these intensity relationships to derive new activity energy expenditure estimation models for thigh and dominant wrist acceleration, by substituting the non-dominant wrist term in our original models with the derived equation to harmonise either dominant wrist or thigh acceleration to non-dominant wrist acceleration.

Activity energy expenditure was estimated separately from each of the acceleration signals by directly applying the appropriate linear and quadratic equations given in Table 2 to 5-second level data; the resulting 5-second level estimated activity energy expenditure signal was then summarised to a mean-per-day average activity energy expenditure using diurnal adjustment to compensate for any between-individual bias introduced by periods of nonwear [15]. To ensure a stable estimate of this circadian model, a minimum of 72 hours of valid data was required per signal to be included in the analyses. Predicted total energy expenditure (in MJ·day$^{-1}$) was calculated as the sum of predicted activity energy expenditure and predicted resting energy expenditure from the simplest model (using only age, sex, height and weight) [40], and dividing the result by 0.9 to account for diet-induced thermogenesis [88]. Agreement between these two predictions against measured activity energy expenditure and total energy expenditure from doubly labelled water was formally tested by calculating the pairwise mean

bias and 95% limits of agreement, Root Mean Squared Error (RMSE) and Pearson's correlation coefficient.

Linear regression was also used to characterise the relationship between the acceleration measurements and activity energy expenditure/total energy expenditure derived from doubly labelled water. As the main focus of this paper is on absolute validity, these relative validity results are supplied in the supplementary material.

The statistical tests were performed using Python v3.6 and Stata v14 (StataCorp, TX, USA).

## 3.3 Results

A descriptive summary of participant characteristics is given in Table 3.1. We recruited 193 participants, and the group measured by doubly labelled water was split equally between men and women. According to the doubly labelled water measurements, mean (standard deviation) total energy expenditure was 11.6 (2.3) MJ·day$^{-1}$, of which 6.6 (1.2) MJ·day$^{-1}$ was resting energy expenditure. Mean (standard deviation) activity-related acceleration (ENMO) per day was 32.4 (8.3) milli-g on the dominant wrist, 28.8 (7.7) milli-g on the non-dominant wrist, and 27.8 (10.9) milli-g on the thigh. Mean dominant wrist acceleration was higher than non-dominant wrist in 84% of participants. Some accelerometry measurements were not included in the analyses due to a combination of devices being lost by participants (n=7), device failures (n=3), files overwritten before download (n=3), and insufficient wear time (n=3). Of those files that overlapped with doubly labelled water measurements, 3 were dominant wrist records, 3 were non-dominant wrist and 9 were thigh records. There was no loss of data in the doubly labelled water, anthropometry or food frequency questionnaire measurements.

| | DLW (n=100) | | | | Non-DLW (n=93) | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Min | Max | Mean | SD | Min | Max |
| Sex (% women) | 50% | | | | 41% | | | |
| Age (years) | 54.4 | 7.2 | 40 | 65 | 54 | 6.7 | 41 | 66 |
| Height (m) | 1.7 | 0.1 | 1.5 | 1.9 | 1.7 | 0.1 | 1.5 | 2.0 |
| Weight (kg) | 78.2 | 13.6 | 48.7 | 110.8 | 77.1 | 12.4 | 56.4 | 112.3 |
| BMI ($kg{\cdot}m^{-2}$) | 26.5 | 3.4 | 20.4 | 36.6 | 25.9 | 2.9 | 20.4 | 35.3 |
| TEE ($MJ{\cdot}day^{-1}$) | 11.6 | 2.3 | 6.5 | 16.4 | - | - | - | - |
| REE ($MJ{\cdot}day^{-1}$) | 6.6 | 1.2 | 3.7 | 9.8 | - | - | - | - |
| DIT fraction | 0.10 | 0.00 | 0.08 | 0.11 | - | - | - | - |
| AEE ($MJ{\cdot}day^{-1}$) | 3.9 | 1.4 | 0.7 | 7.6 | - | - | - | - |
| AEE ($kJ{\cdot}day^{-1}{\cdot}kg^{-1}$) | 49.8 | 16.3 | 8.5 | 92.6 | - | - | - | - |
| $k_O$ ($vSMOW{\cdot}day^{-1}$) | 0.119 | 0.03 | 0.066 | 0.257 | - | - | - | - |
| $k_H$ ($vSMOW{\cdot}day^{-1}$) | 0.093 | 0.028 | 0.044 | 0.228 | - | - | - | - |
| $N_O$ (moles) | 2124 | 434 | 1215 | 3131 | - | - | - | - |
| $N_H$ (moles) | 2188 | 447 | 1251 | 3224 | - | - | - | - |
| DW ENMO (milli-g) | 32.4 | 8.3 | 15.4 | 64.7 | 33.1 | 10.5 | 18.8 | 82.4 |
| NDW ENMO (milli-g) | 28.8 | 7.7 | 15.6 | 59.0 | 29.3 | 8.3 | 16.2 | 63.2 |
| Thigh ENMO (milli-g) | 27.8 | 10.9 | 13.2 | 76.3 | 28.2 | 10.0 | 12.6 | 80.5 |
| DW HPFVM (milli-g) | 48.5 | 11.0 | 25.7 | 85.9 | 49.6 | 12.8 | 31.4 | 105.7 |
| NDW HPFVM (milli-g) | 43.5 | 10.3 | 25.8 | 85.4 | 44.7 | 11.0 | 27.3 | 89.2 |
| Thigh HPFVM (milli-g) | 37.4 | 12.7 | 17.7 | 77.0 | 38.6 | 11.8 | 17.7 | 94.6 |

Key: BMI=Body Mass Index, TEE=Total Energy Expenditure, REE=Resting Energy Expenditure, DIT=Diet-induced Thermogenesis, AEE=Activity Energy Expenditure, DW=Dominant Wrist, NDW=Non-Dominant Wrist $k_O$ and $k_H$=disappearance rates of Oxygen and Hydrogen according to DLW, respectively. $N_O$ and $N_H$=biological pool sizes of Oxygen and Hydrogen, respectively. vSMOW=Vienna-Standard Mean Ocean Water.

Table 3.1: Participant characteristics, provided separately for the doubly labelled water and non-doubly labelled water groups.

Table 3.2 lists the derived equations to predict activity energy expenditure from each of the sensors, as informed by the harmonisation equations which are supplied in Supplementary Table 3.5. For brevity, Table 3.3 summarises the absolute validity of the

quadratic HPFVM models applied to measurements from both wrists and thigh with respect to activity energy expenditure, and Table 3.3 summarises agreement with total energy expenditure derived from doubly labelled water. Bland-Altman plots illustrating the agreement of these estimates is supplied in Figures 3.1 and 3.2. A table summarising the remaining models is given in Supplementary Table 3.9.

| Placement | Metric | Formulae to estimate AEE in J·min$^{-1}$·kg$^{-1}$ |
|---|---|---|
| NDW | ENMO | $5.01 + 1.000 \times x$ |
| NDW | ENMO | $-10.58 + 1.1176 \times x$ $+ 2.9418 \times sqrt(x) - 0.00059277 \times (x^2)$ |
| NDW | HPFVM | $-4.65 + 0.8537 \times x$ |
| NDW | HPFVM | $-1.25 + 1.1353 \times x$ $- 2.4281 \times sqrt(x) - 0.00040270 \times (x^2)$ |
| DW | ENMO | $5.01 + 1.000 \times (1.5 + .8517 \times x)$ |
| DW | ENMO | $-10.58 + 1.1176 \times (1.5 + .8517 \times x)$ $+ 2.9418 \times sqrt((1.5 + .8517 \times x))$ $- 0.00059277 \times ((1.5 + .8517 \times x)^2)$ |
| DW | HPFVM | $-4.65 + 0.8537 \times (1.3 + .8781 \times x)$ |
| DW | HPFVM | $-1.25 + 1.1353 \times (1.3 + .8781 \times x)$ $- 2.4281 \times sqrt((1.3 + .8781 \times x))$ $- 0.00040270 \times ((1.3 + .8781 \times x)^2)$ |
| Thigh | ENMO | $5.01 + 1.000 \times (13.4 + .5674 \times x)$ |
| Thigh | ENMO | $-10.58 + 1.1176 \times (13.4 + .5674 \times x)$ $+ 2.9418 \times sqrt((13.4 + .5674 \times x))$ $- 0.00059277 \times ((13.4 + .5674 \times x)^2)$ |
| Thigh | HPFVM | $-4.65 + .8537 \times (20.3 + .6401 \times x)$ |
| Thigh | HPFVM | $-1.25 + 1.1353 \times (20.3 + .6401 \times x)$ $- 2.4281 \times sqrt((20.3 + .6401 \times x))$ $- 0.00040270 \times ((20.3 + .6401 \times x)^2)$ |

Table 3.2: Derived linear and quadratic equations to estimate activity energy expenditure (J·min$^{-1}$·kg$^{-1}$) from wrist and thigh acceleration intensity. (4.184 J·min$^{-1}$·kg$^{-1}$ = 1 cal·min$^{-1}$·kg$^{-1}$, and 71.225 J·min$^{-1}$·kg$^{-1}$ = 1 net Metabolic Equivalent Task (MET)).

| Placement | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|
| Activity energy expenditure [via primary REE] | | | | | | |
| Dominant wrist | 97 | -1.9 | -26.0 | 22.2 | 0.644 | 12.4 |
| Non-dominant wrist | 97 | -1.5 | -25.1 | 22.1 | 0.676 | 12.1 |
| Thigh | 91 | -4.2* | -29.6 | 21.2 | 0.599 | 13.6 |
| Both wrists | 94 | -1.9 | -25.1 | 21.3 | 0.669 | 11.9 |
| Non-dominant wrist & Thigh | 89 | -3.3 | -26.2 | 19.6 | 0.687 | 12.1 |
| Dominant wrist & Thigh | 88 | -3.5 | -27.2 | 20.1 | 0.644 | 12.5 |
| Both wrists & Thigh | 86 | -3.4 | -25.9 | 19.2 | 0.675 | 11.9 |
| Activity energy expenditure [via measured REE only] | | | | | | |
| Dominant wrist | 97 | -0.5 | -26.1 | 25.2 | 0.613 | 13.0 |
| Non-dominant wrist | 97 | -0.2 | -25.1 | 24.7 | 0.649 | 12.6 |
| Thigh | 91 | -3.0 | -29.8 | 23.9 | 0.570 | 13.9 |
| Both wrists | 94 | -0.6 | -25.0 | 23.9 | 0.644 | 12.4 |
| Non-dominant wrist & Thigh | 89 | -2.2 | -26.3 | 21.9 | 0.661 | 12.4 |
| Dominant wrist & Thigh | 88 | -2.3 | -27.5 | 22.9 | 0.610 | 13.0 |
| Both wrists & Thigh | 86 | -2.2 | -26.0 | 21.5 | 0.650 | 12.2 |

Table 3.3: Agreement between estimated activity energy expenditure from the HPFVM quadratic models with those derived from doubly labelled water. An asterisk (*) next to a bias value indicates statistical significance according to a paired t-test ($p < 0.05$).

| Placement | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|
| Total energy expenditure | | | | | | |
| Dominant wrist | 97 | -0.3 | -2.2 | 1.7 | 0.903 | 1.0 |
| Non-dominant wrist | 97 | -0.3 | -2.1 | 1.6 | 0.911 | 1.0 |
| Thigh | 91 | -0.5 | -2.7 | 1.7 | 0.874 | 1.2 |
| Both wrists | 94 | -0.3 | -2.1 | 1.6 | 0.911 | 1.0 |
| Non-dominant wrist & Thigh | 89 | -0.4 | -2.3 | 1.5 | 0.909 | 1.0 |
| Dominant wrist & Thigh | 88 | -0.4 | -2.4 | 1.5 | 0.902 | 1.1 |
| Both wrists & Thigh | 86 | -0.4 | -2.2 | 1.4 | 0.914 | 1.0 |

Table 3.4: Agreement between estimated total energy expenditure from the HPFVM quadratic models with those derived from doubly labelled water. An asterisk (*) next to a bias value indicates statistical significance according to a paired t-test ($p < 0.05$).

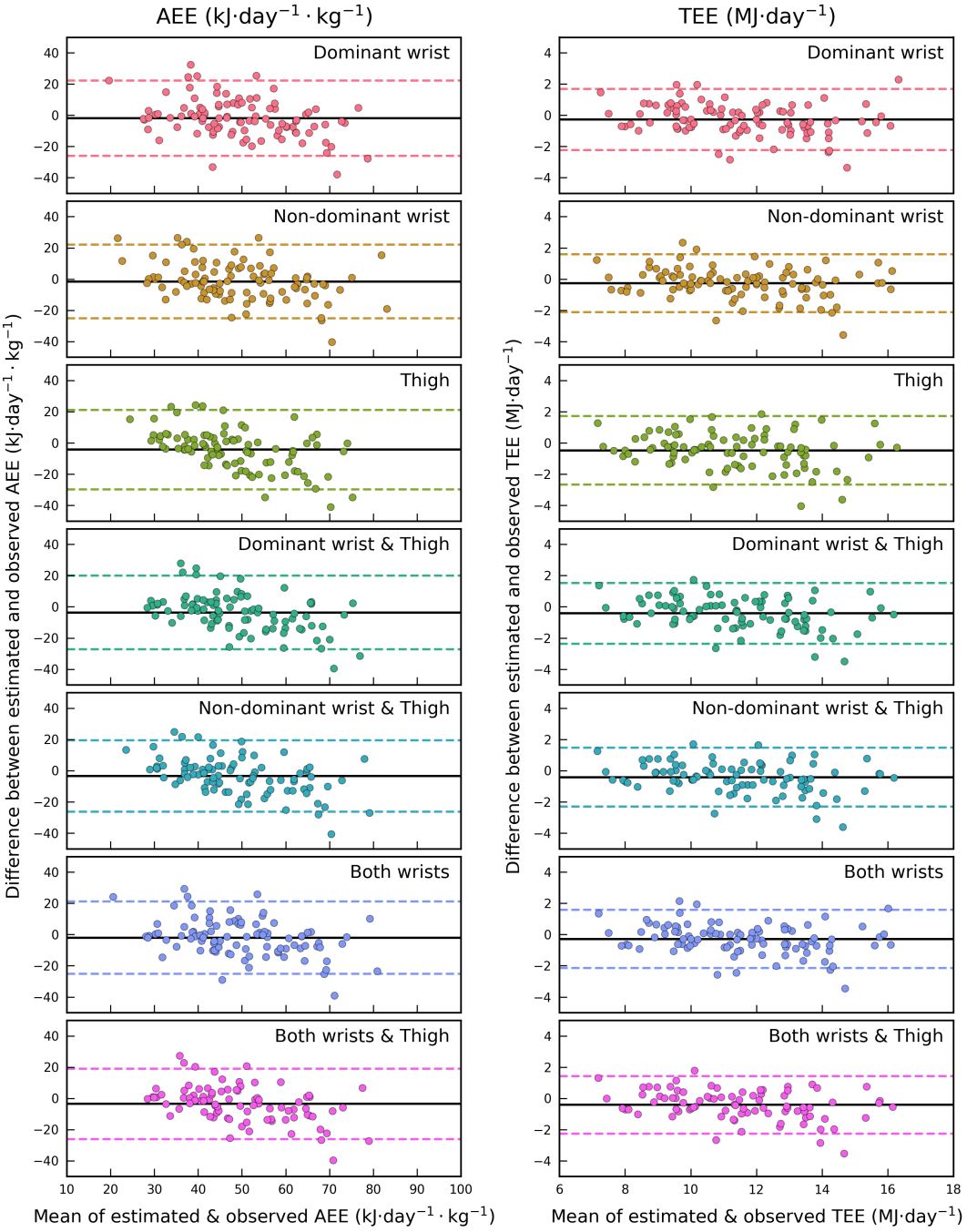


Figure 3.1: Bland-Altman plots illustrating agreement between the activity energy expenditure and total energy expenditure estimates from HPFVM Quadratic models with those from doubly labelled water, where the X-axis indicates the mean of measured and observed values.

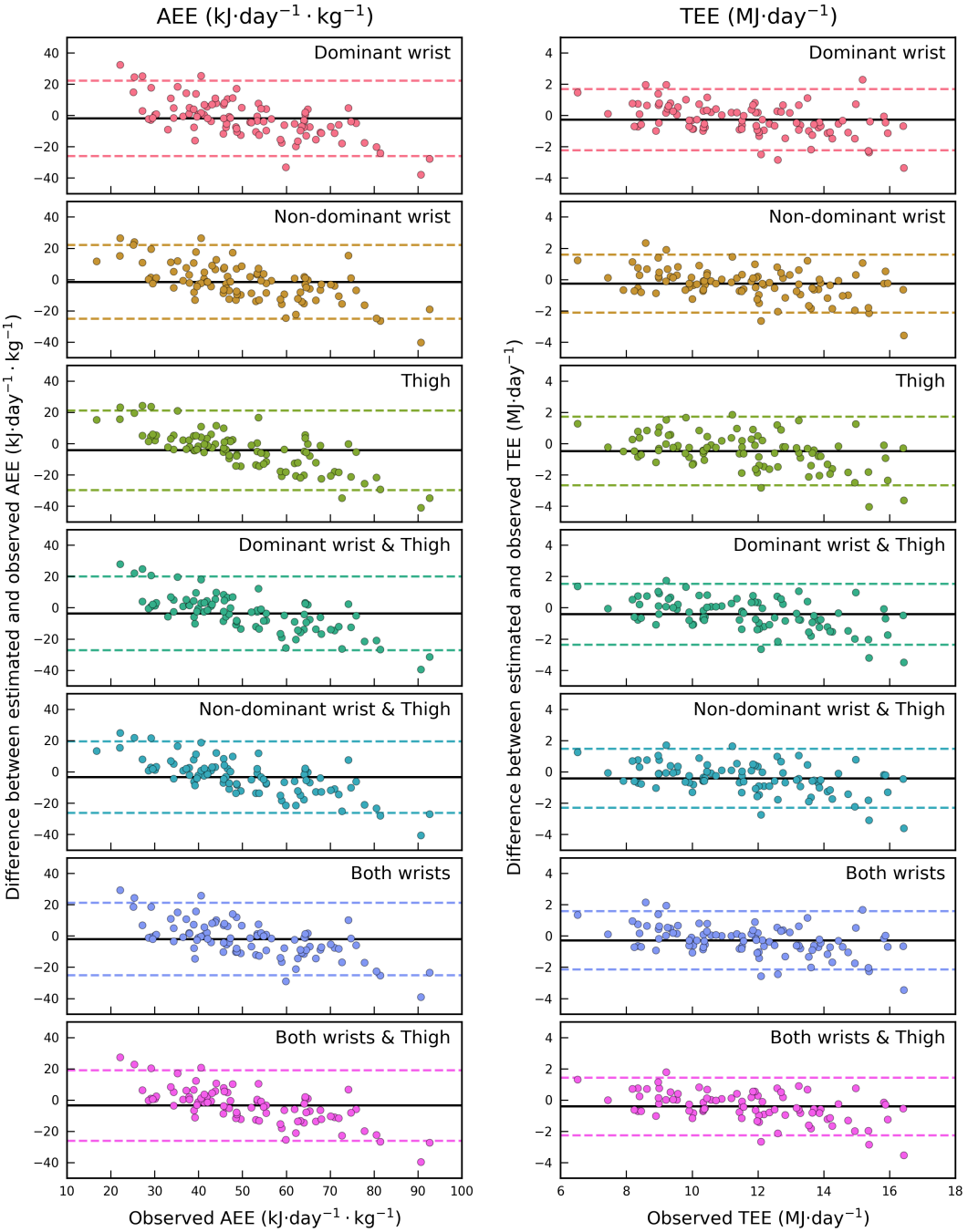


Figure 3.2: Bland-Altman plots illustrating agreement between the activity energy expenditure and total energy expenditure estimates from HPFVM Quadratic models with those from doubly labelled water, where the X-axis indicates the gold-standard observed value.

The difference in performance between each estimation model was very minor; all activity energy expenditure estimates had small negative mean biases (underestimates) at the population level (average -2.8 kJ·day$^{-1}$·kg$^{-1}$) but of these only the thigh model biases were statistically significant. RMSEs for activity energy expenditure ranged from 11.9 to 13.5 kJ·day$^{-1}$·kg$^{-1}$ (24 to 27% of the mean), and 1.0 to 1.2 MJ·day$^{-1}$ for total energy expenditure (8 to 10% of the mean). Pearson correlations ranged from 0.6 to 0.69 with activity energy expenditure, and from 0.87 to 0.91 with total energy expenditure. Combined estimates using two or more sensors lead to very negligible performance improvements over single-sensor estimates. Signed estimation errors were nominally positively correlated with body fat percentage when using our primary characterisation of resting energy expenditure (r=0.18-0.25), and less so with exclusively measured values (r=0.10-0.17).

In the non-doubly labelled water group, 88 participants had at least 3 days of valid simultaneous wrist signals during free-living, and 84 had simultaneous wrist and thigh signals; around 200,000 5-minute observations included in each of the regression analyses. The between-individual explained variance between dominant and non-dominant wrist intensity signals was approximately 86% (99% within-individual), and the average between-individual explained variance between wrist and thigh intensities was approximately 49% (97% within-individual). The derived linear models to harmonise the acceleration signals are listed in Supplementary Table 3.5. The final models given to estimate activity energy expenditure from dominant wrist and thigh in Table 2 were the result of substituting these harmonisation equations into the original non-dominant wrist models.

## 3.4 Discussion

In this work, we have applied our previously derived activity intensity estimation models [91] to wrist acceleration signals (after harmonising the intensity of dominant wrist to non-dominant wrist) and investigated their agreement with a gold-standard measure of activity energy expenditure. We arrived at estimates that were highly correlated with the criterion ($r > 0.6$) with small and non-significant mean biases at the population level from both wrists and low RMSEs of approximately 12 kJ·day$^{-1}$·kg$^{-1}$. We have also introduced and validated new intensity estimation models for thigh acceleration, demonstrating similar performance to the wrist models. We observed that dominant wrist acceleration was on average 12% higher than non-dominant wrist in free-living individuals, but that those measures were very highly correlated (r=0.93), allowing us to derive conversion models which harmonise acceleration intensity measured at either wrist. To our knowledge, this is the first demonstration of the absolute validity of a time-integrated predictive model of activity intensity for either wrist or thigh accelerometry.

Our findings on the high correlation between dominant wrist and non-dominant wrist acceleration in free-living individuals are consistent with a previous study in a small convenience sample (n=40) [24]. They observed 5% higher dominant wrist acceleration compared to non-dominant wrist, but it was not a statistically significant difference, perhaps due to the shorter duration of measurement and smaller sample size. In our relative validity tests, we found that each wrist separately explained a similar variance in activity energy expenditure, and inclusion of both wrist measurements in the linear models did not drastically improve performance over either wrist measurement alone. Taken together, these results are indicative of a high degree of upper-body symmetry. One implication of these findings is that irrespective of hand dominance, wrist acceler-

ation measurements are naturally conducive to harmonisation across studies, making them well suited to pooled- and meta-analysis. Conversely, it implies that implementing dual wrist measurements may be a largely redundant exercise for studies whose primary intention is to capture activity energy expenditure. However, there is a possibility that future methodological advances in the field of activity recognition may be able to better utilise simultaneous wrist signals, which could yield a more precise estimation of instantaneous activity energy expenditure.

The estimation models validated herein for the wrist were derived using a training dataset in which non-dominant wrist acceleration data was collected at 60 Hz with a GeneActiv device [91], and were successfully validated using 100 Hz data collected with an Axivity AX3. With an additional harmonisation step, the model also translated to acceptably strong inferences on the dominant wrist, albeit with a slightly increased error. This indicates that our models capture a generalized biomechanical relationship of wrist movement, rather than being superficial transformations of a specific device's output to activity energy expenditure. It therefore suggests that these models are applicable to any wrist-worn device which provides raw, unfiltered triaxial acceleration data expressed in SI units.

The associations between wrist acceleration and observations from DLW have been reported before, in pregnant and non-pregnant Swedish women [83]. In that population it explained 27% of the variance in activity energy expenditure (kJ·day$^{-1}$·kg$^{-1}$) in non-pregnant women (n=48), but only 5% in pregnant women (n=26); however, those wrist measurements were evenly divided between left and right wrist, which most likely lead to a mix of dominant and non-dominant wrist measurements and potentially attenuated the correlations.

The previously established estimation models applied to the non-dominant wrist resulted in robust estimates with small, non-significant mean biases, which is a strong

justification for using this inference scheme to infer activity energy expenditure in free-living individuals. The higher average of the dominant wrist would have led to a significant overestimation had we applied the original non-dominant wrist model, but our harmonisation approach effectively scaled the dominant wrist measure down to the level of non-dominant wrist, ultimately leading to virtually identical results. We note that physical activity was measured by dominant wrist accelerometry in UK Biobank [25]. We have now demonstrated the validity of this approach in a demographically comparable sample of UK adults. Specifically, the absolute validity result for ENMO in Supplementary Table 2 demonstrates that our linear estimation model applied to ENMO at 5-second resolution yielded a valid activity energy expenditure estimate, with a small mean bias and a RMSE of 13 kJ·day$^{-1}$·kg$^{-1}$ and high correlation (r=0.61). Consequently, we can use the equations for dominant wrist in Table 2 to solve for specific energy expenditure values – for example, 3 metabolic equivalents (activity energy expenditure 142 J·min$^{-1}$·kg$^{-1}$) is the generally accepted threshold for "moderate" activity intensity, and our ENMO equation suggests this is approximately 159 milli-g on the dominant wrist.

Our findings for the thigh acceleration models demonstrate that thigh-worn accelerometers capture an information-rich biomechanical signal, from which valid estimates of activity energy expenditure can be made. As a consequence of the larger y-intercepts of the thigh models, their minimum estimated activity energy expenditure ranges from 10 to 18 J·min$^{-1}$·kg$^{-1}$ (0.15-0.25 metabolic equivalents). To our knowledge, only one previous study has described the association between thigh acceleration and activity energy expenditure from doubly labelled water, in a small study of free-living cancer patients and controls [74]; which reported very low agreement between the manufacturer's proprietary activity energy expenditure prediction and the criterion. While thigh-worn sensors do not yet have the same popularity as wrist-worn sensors [73, 79], large-scale data collections are planned for the future [41]. Our models enable new analyses to be

conducted in those existing datasets, and may make thigh-worn accelerometry a more appealing option for future studies if issues of feasibility can be addressed.

The results of our absolute validity tests demonstrate that deriving intensity models using a "silver-standard" criterion (such as individually-calibrated heart rate and uniaxial movement sensing) in a large sample of free-living adults is a sound approach. The combined sensing estimate of activity energy expenditure is less precise than respiratory gas analysis which can be captured in laboratory studies [12] but there are several reasons why we have been able to derive superior models to previous approaches. Firstly, the dataset was collected in free-living participants, and is therefore representative of the intended application, as opposed to artificial scenarios and activities performed in a laboratory. Secondly, the combined sensing approach embedded in a cohort study allowed the collection of a volume of data many orders of magnitude greater than any laboratory study has for this purpose. Our training dataset alone contained over 16.6 person-years of observation ($>$1.7 million data points). One disadvantage of this approach is that we are unable to capture categorical labelled data, so there is no opportunity to explore activity type recognition.

It is appropriate to compare our absolute validity results here with those of combined sensing itself [14]. In our previous work, the best estimate with treadmill test calibration resulted in a RMSE of 20 kJ·day$^{-1}$·kg$^{-1}$ (30% of the 66 kJ·day$^{-1}$·kg$^{-1}$ criterion mean), non-significant positive mean bias of approximately 4 kJ·day$^{-1}$·kg$^{-1}$ (6%) at the population level, and a correlation of 0.67 in a sample of 50 UK adults. Compared to the present results, all estimations here had considerably lower RMSEs of around 12 kJ·day$^{-1}$·kg$^{-1}$ (25% of the 50 kJ·day$^{-1}$·kg$^{-1}$ mean), similar magnitude but negative mean biases ( 6%), and slightly higher correlations. However, our study participants were significantly less active overall according to the criterion, ultimately leading to a similar relative accuracy. Combined sensing model errors were also uncorrelated to

body fat percentage, whereas errors of accelerometry-only models seem to display this characteristic, albeit less so in the present study (r=0.22 versus r=0.63 for uniaxial trunk acceleration). Contrasting the feasibility of the methods, however, wrist accelerometry has the advantages of being cheaper, less burdensome to both participants and research staff, and does not require individual calibration using an exercise test. Comparing performance of other devices worn on the upper limbs, validation of the now-discontinued SenseWear Pro3 and Mini also achieved no significant bias with respect to total energy expenditure, but with lower correlations (r=0.84) than any of our total energy expenditure models (r=0.9) and wider limits of agreement [45] and with lower feasibility.

Some have suggested that simple movement intensity approaches should be replaced by more sophisticated models that utilise a broader range of signal features [28, 57]. Recent efforts to estimate energy expenditure have utilised a range of machine learning approaches, such as neural networks [56, 52, 58] and random forests [28]. While we are not aware of any such methodology with a performance that exceeds the simpler models validated in this paper, this is an interesting area of future work.

In summary, we have evaluated the absolute validity of intensity models of activity energy expenditure from wrist and thigh accelerometry, and concluded that they provide precise and accurate estimates in free-living adults. With the addition of predicted resting energy expenditure to produce total energy expenditure, we found even stronger validity at the population level. Considering its feasibility, wrist accelerometry emerges as a viable candidate for deployment in a large scale studies, including physical activity surveillance and the prediction of total energy expenditure in dietary surveys.

## 3.5 Acknowledgements

We are very grateful to the participants who took part in this study. We thank the principal investigators of the Fenland study for allowing us to recruit from this study population, and the functional teams of the MRC Epidemiology Unit (Study Coordination, Field Epidemiology, Anthropometry, Data Management, IT) for supporting the study. We would like to specifically acknowledge Lewis Griffiths, Katie Palmer, and Eoin McNamara for their assistance in the data collection for this study, and Annie Schiff, Richard Salisbury and Nicola Kimber for study co-ordination and recruitment.

We would like to thank Eirini Trichia from the MRC Epidemiology Unit for processing the FFQ data with the FETA package. We would also like to thank the stable isotope team from the MRC Elsie Widdowson Laboratory: Priya Singh, Elise Orford and Kevin Donkers for the DLW preparation and analysis.

Medical Research Council and UK Biobank are acknowledged for covering costs of the fieldwork. Newcastle University and MedImmune are acknowledged for covering the costs of the doubly labelled water measurements.

## 3.6 Supplementary material

| Formula | RMSE | Within $r^2$ | Between $r^2$ | Obs. | N |
|---|---|---|---|---|---|
| ENMO | | | | | |
| NDW = 1.5 + .8516909× DW | 17.3 | 0.887 | 0.867 | 209,537 | 88 |
| NDW = 13.4 + .5674314× T | 28.4 | 0.644 | 0.537 | 199,170 | 84 |
| DW = 2.2 + 1.041485× NDW | 19.1 | 0.887 | 0.867 | 209,537 | 88 |
| DW = 16.3 + .5898457× T | 34.7 | 0.569 | 0.429 | 204,043 | 88 |
| T = -3.0 + .9650111× DW | 44.4 | 0.569 | 0.429 | 204,043 | 88 |
| T = -4.8 + 1.135703× NDW | 40.2 | 0.644 | 0.537 | 199,170 | 84 |
| HPFVM | | | | | |
| NDW = 1.3 + .8781101× DW | 19.4 | 0.909 | 0.864 | 209,537 | 88 |
| NDW = 20.3 + .6401075× T | 36.1 | 0.666 | 0.538 | 199,170 | 84 |
| DW = 3.1 + 1.035547× NDW | 21.1 | 0.909 | 0.864 | 209,537 | 88 |
| DW = 24.7 + .6491317× T | 43.4 | 0.588 | 0.434 | 204,043 | 88 |
| T = -5.9 + .9067731× DW | 51.3 | 0.588 | 0.434 | 204,043 | 88 |
| T = -7.8 + 1.0416× NDW | 46.0 | 0.666 | 0.538 | 199,170 | 84 |

Table 3.5: Derived equations to harmonise acceleration intensity measured at the dominant wrist, non-dominant wrist and thigh. Formulae and RMSE expressed in milli-g.

| | Mean | SD | Min | Max |
|---|---|---|---|---|
| TEE (MJ·day$^{-1}$) | 11.6 | 2.3 | 6.5 | 16.4 |
| REE (MJ·day$^{-1}$) [primary] | 6.6 | 1.2 | 3.7 | 9.9 |
| REE (MJ·day$^{-1}$) [estimated only] | 6.4 | 1.0 | 4.6 | 8.9 |
| REE (MJ·day$^{-1}$) [measured only] | 6.7 | 1.4 | 3.7 | 10.0 |
| DIT fraction | 0.097 | 0.008 | 0.082 | 0.118 |
| AEE (kJ·day$^{-1}$·kg$^{-1}$) [primary REE] | 49.8 | 16.3 | 8.5 | 92.6 |
| AEE (kJ·day$^{-1}$·kg$^{-1}$) [estimated REE only] | 52.3 | 16.8 | 10.9 | 93.3 |
| AEE (kJ·day$^{-1}$·kg$^{-1}$) [measured REE only] | 48.4 | 16.8 | 3.9 | 99.2 |

Table 3.6: Resting energy expenditure summaries according to the different characterisastions, and the consequent activity energy expenditure summaries in the doubly labelled water sample (n=100).

| Formula | $r^2$ | RMSE | N |
|---|---|---|---|
| HPFVM | | | |
| $y = 4.656 + 0.945 \times DW$ | 0.42 | 101.53 | 97 |
| $y = 3.649 + 1.068 \times NDW$ | 0.45 | 107.37 | 97 |
| $y = 21.945 + 0.755 \times T$ | 0.35 | 90.95 | 91 |
| $y = 3.401 + 0.412 \times DW + 0.624 \times NDW$ | 0.45 | 72.40 | 94 |
| $y = 8.307 + 0.718 \times DW + 0.208 \times T$ | 0.42 | 67.44 | 88 |
| $y = 6.173 + 0.871 \times NDW + 0.177 \times T$ | 0.48 | 73.53 | 89 |
| $y = 7.337 + 0.715 \times NDW + 0.146 \times DW + 0.148 \times T$ | 0.46 | 56.27 | 86 |
| ENMO | | | |
| $y = 12.641 + 1.167 \times DW$ | 0.37 | 95.41 | 97 |
| $y = 11.335 + 1.347 \times NDW$ | 0.40 | 101.12 | 97 |
| $y = 27.619 + 0.813 \times T$ | 0.30 | 83.91 | 91 |
| $y = 10.830 + 0.526 \times DW + 0.789 \times NDW$ | 0.41 | 68.99 | 94 |
| $y = 15.415 + 0.907 \times DW + 0.215 \times T$ | 0.37 | 63.27 | 88 |
| $y = 12.613 + 1.182 \times NDW + 0.140 \times T$ | 0.43 | 69.66 | 89 |
| $y = 13.093 + 0.946 \times NDW + 0.277 \times DW + 0.061 \times T$ | 0.42 | 53.68 | 86 |

Table 3.7: Observed linear relationships between AEE (normalised for body weight) and summarised acceleration measures.

| Formula | $r^2$ | RMSE | N |
|---|---|---|---|
| HPFVM | | | |
| $y = -1846.7 + 57.6 \times DW + 0.216 \times (DW \times weight)$ | 0.505 | 5466.5 | 97 |
| $y = 288.1 + 19.3 \times NDW + 0.817 \times (NDW \times weight)$ | 0.512 | 5528.4 | 97 |
| $y = -727.9 + 60.5 \times T + 0.008 \times (T \times weight)$ | 0.441 | 4927.6 | 91 |
| ENMO | | | |
| $y = -1821.4 + 88.5 \times DW + 0.045 \times (DW \times weight)$ | 0.457 | 5198.5 | 97 |
| $y = 289.2 + 29.1 \times NDW + 0.962 \times (NDW \times weight)$ | 0.460 | 5242.2 | 97 |
| $y = -1924.1 + 133.2 \times T + -0.809 \times (T \times weight)$ | 0.406 | 4725.3 | 91 |

Table 3.8: Observed linear relationships between absolute AEE (not normalised for body weight) and summarised acceleration measures, with interactions on body weight.

| Placement | Metric & model | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|---|
| DW | ENMO Linear | 97 | -1.4 | -26.2 | 23.4 | 0.610 | 12.7 |
| DW | ENMO Quadratic | 97 | -3.6 | -28.3 | 21.2 | 0.615 | 13.1 |
| DW | HPFVM Linear | 97 | -2.4 | -26.4 | 21.5 | 0.648 | 12.4 |
| DW | HPFVM Quadratic | 97 | -1.9 | -26.0 | 22.2 | 0.644 | 12.4 |
| NDW | ENMO Linear | 97 | -1.3 | -26.1 | 23.5 | 0.631 | 12.6 |
| NDW | ENMO Quadratic | 97 | -3.3 | -27.9 | 21.4 | 0.640 | 12.9 |
| NDW | HPFVM Linear | 97 | -2.0 | -25.8 | 21.7 | 0.673 | 12.2 |
| NDW | HPFVM Quadratic | 97 | -1.5 | -25.1 | 22.1 | 0.676 | 12.1 |
| Thigh | ENMO Linear | 91 | -1.0 | -27.6 | 25.6 | 0.546 | 13.5 |
| Thigh | ENMO Quadratic | 91 | -1.0 | -26.6 | 24.6 | 0.591 | 13.0 |
| Thigh | HPFVM Linear | 91 | -2.5 | -28.1 | 23.1 | 0.592 | 13.2 |
| Thigh | HPFVM Quadratic | 91 | -4.2* | -29.6 | 21.2 | 0.599 | 13.6 |
| DW & T | ENMO Linear | 88 | -1.6 | -26.3 | 23.1 | 0.601 | 12.6 |
| DW & T | ENMO Quadratic | 88 | -2.7 | -26.8 | 21.4 | 0.627 | 12.5 |
| DW & T | HPFVM Linear | 88 | -3.0 | -26.6 | 20.7 | 0.643 | 12.4 |
| DW & T | HPFVM Quadratic | 88 | -3.5 | -27.2 | 20.1 | 0.644 | 12.5 |
| NDW & T | ENMO Linear | 89 | -1.5 | -25.8 | 22.8 | 0.638 | 12.4 |
| NDW & T | ENMO Quadratic | 89 | -2.5 | -26.0 | 20.9 | 0.670 | 12.2 |
| NDW & T | HPFVM Linear | 89 | -2.7 | -25.8 | 20.3 | 0.681 | 12.0 |
| NDW & T | HPFVM Quadratic | 89 | -3.3 | -26.2 | 19.6 | 0.687 | 12.1 |
| BW | ENMO Linear | 94 | -1.6 | -25.6 | 22.4 | 0.635 | 12.3 |
| BW | ENMO Quadratic | 94 | -3.7 | -27.4 | 20.0 | 0.649 | 12.6 |
| BW | HPFVM Linear | 94 | -2.4 | -25.7 | 20.8 | 0.667 | 12.1 |
| BW | HPFVM Quadratic | 94 | -1.9 | -25.1 | 21.3 | 0.669 | 11.9 |
| BW & T | ENMO Linear | 86 | -2.0 | -25.6 | 21.7 | 0.634 | 12.2 |
| BW & T | ENMO Quadratic | 86 | -3.4 | -26.4 | 19.6 | 0.661 | 12.1 |
| BW & T | HPFVM Linear | 86 | -3.1 | -25.8 | 19.5 | 0.672 | 11.9 |
| BW & T | HPFVM Quadratic | 86 | -3.4 | -25.9 | 19.2 | 0.675 | 11.9 |

Table 3.9: Agreement between estimated AEE from all models with those derived from DLW.

Chapter 4

# Deep convolutional neural networks to estimate activity energy expenditure from wearable sensors

*Introduction: In many large studies worldwide, the physical activity of participants during free-living has been recorded using body-worn sensors which log raw triaxial acceleration. Traditional methods estimate physical activity energy expenditure (PAEE) from this data by calculating movement intensity and applying simple equations. It has been hypothesised that these estimates could be improved upon by better utilising more features of the signal.*

*Methods: 193 UK adults wore an accelerometer on each wrist and right thigh for up to 9 days during free-living. They simultaneously wore a heart-rate and movement sensor, which was interpreted using an exercise test (performed at a clinic beforehand) to produce an individually-calibrated PAEE signal. 100 of the participants were also measured by doubly labelled water (DLW) to assess total energy expenditure (TEE). Deep convolutional neural networks were trained to estimate the PAEE signal from each of the separate raw acceleration signals in the non-DLW group. These models were evaluated in the DLW group, by assessing agreement with AEE derived from TEE.*

*Results: Mean TEE and AEE derived from DLW was 11.6 (2.3) MJ·day$^{-1}$ and 49.8 (16.3) kJ·day$^{-1}$·kg$^{-1}$. There were small positive mean estimation biases at the population-level, of which only the dominant wrist was statistically significant (4.2 kJ·day$^{-1}$·kg$^{-1}$, p=0.035). All estimates correlated highly with the criterion (average r=0.7) and root mean squared errors averaged 11.9 kJ·day$^{-1}$·kg$^{-1}$ (24% error).*

*Discussion: The performance of each deep neural network was an improvement upon its movement intensity counterpart. Neural networks have great potential for the interpretation of wearable sensor data.*

## 4.1 Introduction

There is a growing worldwide obesity epidemic, and an increasing burden of related metabolic diseases like type 2 diabetes. In order to understand the causes of these trends, there is an urgent need for tools to accurately assess energy intake and energy expenditure [34], which together contribute to energy homeostasis. Activity energy expenditure refers to energy expended by the movement of skeletal muscle [18]; it is the most variable component of total energy expenditure and has proven difficult to assess in free-living conditions [86]. Wearable sensors such as accelerometers are often employed in large scale epidemiological studies to capture free-living physical activity [25, 78, 22], and wrist-worn devices in particular have become the most popular form-factor due to their low burden upon the participant [83, 66], which has lead to greatly improved protocol adherence and the preservation of study sample sizes for further analyses. The newest such devices collect raw acceleration signals in three dimensions at a high frequency, providing a high resolution record of movements. New methodologies are required to utilise the information contained in these signals in order to optimally estimate activity energy expenditure for an individual.

The conventional approach to estimating activity energy expenditure from accelerometry measures is to characterise the biomechanical relationship between the intensity of the measured movement and its associated energy cost [64, 70], using linear models or similar parametric approaches. We have already pursued this approach in previous work to derive movement intensity based models for the non-dominant wrist [91]. We then later demonstrated the absolute validity of these models, with results being non-biased when compared against a gold-standard measure of total energy expenditure in free-living, doubly labelled water. To the best of our knowledge, these models currently represent our best approach to estimating activity energy expenditure from accelerome-

ter recordings [64], but there is still much room for improvement in estimation precision. Movement intensity based inferences rely upon the assumption that the measured body part is consistently representative of whole-body movement. For virtually any anatomical site, one can think of an activity in which its specific movement intensity is likely to be misleading, the most commonly cited examples being wrist or hip movement during cycling. This is problematic because the estimates are then affected by the prevalence of those activities where representativeness is an issue, making comparisons across cultures and countries less reliable.

Figure 4.1 shows several example traces measured at the non-dominant wrist during activities such as walking, cycling and hand washing. It is evident that the rich signal collected by an accelerometer contains more information than just average movement intensity; there are clear patterns and repetitive structures that have the potential to be informative when estimating the associated energy cost of the movement. However, utilising such features in complex data requires more sophisticated modelling techniques which are capable of capturing highly non-linear relationships.


Figure 4.1: An array of example wrist acceleration traces during walking, cycling and hand washing.

Deep learning is a term applied to recent methodological advancements in machine learning, referring most commonly to neural network models which work by constructing deep, hierarchical representations of the raw input data [49]. There are many domains where these deep models significantly outperform “classical” modelling approaches, which typically rely heavily on pre-processing the input data, and extracting predetermined features and representations that are suspected to be useful. One such task where deep learning has clearly surpassed traditional approaches is human activity recognition [35, 33, 65]; the task of finding the correct categorical label for a given sequence of data recorded by wearable sensors such as accelerometers. For example, in the OPPORTUNITY [67] locomotion recognition challenge, traditional methods achieved $F_1$ scores of between 0.64-0.87 [19], and deep neural networks have since

reached at least 0.93 [62].

Neural networks designed to make inferences from wearable sensor data typically feature several common topological elements [65]. Firstly, convolutional layers perform matrix convolution operations upon the raw input data along the time axis using small kernels, where each convolution operation derives a feature for the next layer; informally, this can be seen as scanning across the input signal for similarity to a short pattern or shape. Max-pooling layers often follow convolutional layers, which perform a data reduction operation by taking the maximum activation of neurons in the previous layer over a local temporal region; this makes the network robust to small translations in time, meaning a feature or pattern can be present in different parts of the input domain without affecting the corresponding estimate. Since the data captured by wearable sensors is inherently a time-series, it is important that models can properly capture dependencies along the time axis; Long Short Term Memory (LSTM) layers [43] are a tractable way to model such temporal relationships, and there is evidence to suggest they are superior for human activity recognition [35]. In order to interpret these time features and perform the final inference, traditional fully-connected layers are typically found at the tail-end of the neural network, which consist of weighted connections from every neuron in the previous layer to every neuron in its layer.

In this study, we describe the application of deep neural networks to the *regression* problem of predicting activity energy expenditure from raw triaxial acceleration measured at several anatomical sites: the dominant wrist, non-dominant wrist and thigh. We use a dataset of acceleration signals collected simultaneously with a silver-standard physical activity energy expenditure signal in free-living conditions to learn predictive models at a fifteen-second resolution. We then evaluate the resulting models in a large, independent sample measured by a gold-standard measure of energy expenditure in free-living, doubly labelled water, and compare their performances against our previ-

ously established movement intensity based models.

## 4.2 Methods

### 4.2.1 Data collection

Participants were recruited from the Fenland Study, a large cohort established to determine the genetic, behavioural and environmental determinants of obesity and type 2 diabetes [61]. The details of this sub-study can be found elsewhere, wherein we also reported the validity of our movement intensity-based approach (chapter 3 and [89]) to estimating activity energy expenditure. The study was carried out in accordance with the Declaration of Helsinki, approved by the local research ethics board, and all participants provided written informed consent.

Participants attended two clinic visits separated by between 9 to 14 days of free-living. During their first clinic visit, participants performed a graded sub-maximal exercise test on a treadmill, to individually calibrate their heart rate response to energy expenditure levels. During the free-living period, they wore a sensor (Actiheart, camNtech) which recorded their heart rate and trunk acceleration once every fifteen seconds, hereafter referred to as a combined sensor. The data collected during the exercise test were then used to interpret the free-living signal to produce an individually-calibrated energy expenditure signal, the methodology for which has been extensively described and validated elsewhere [75, 12, 13, 14].

Participants were also fitted with accelerometers setup to record raw, triaxial acceleration at 100 Hertz with a dynamic range of $\pm$ 8 g (AX3, Axivity). One accelerometer was worn on each wrist in a silicone wrist band, and one was attached to the right thigh by an adhesive medical dressing. The participants were advised that the devices are wa-

terproof, and were asked to wear the devices continuously for the following eight days and nights, including during showering and bathing. The exact instructions given to participants regarding monitor placement and position are provided in the supplementary material.

Weight was measured to the nearest 0.1 kg using calibrated digital scales (TANITA model BC-418 MA; Tanita, Tokyo, Japan) at both visits. Height was measured to the nearest 0.1 cm using a stadiometer (SECA 240; Seca, Birmingham, UK) at the first clinic visit. Body composition was also measured by DXA (Lunar Prodigy Advanced, GE Healthcare, USA) as part of the Fenland study.

One hundred participants were additionally measured by doubly labelled water, a gold-standard measure of total energy expenditure in free-living [72, 71]. An appropriate individualised $_2H^{18}O$ dosage was determined from self-reported weight (70 mg $^2H_2O$ and 174 mg $H_2{}^{18}O$ per kg body weight), prior to the first clinic visit. Two baseline urine samples were taken prior to dosing, one of which was brought from home and one was taken at the clinic. The participants were then asked to take one urine sample per day until their next clinic visit, preferably at approximately the same time of day and not the first void after waking. Isotopic enrichment was measured by isotope ratio mass spectrometry ($^2$H, Isoprime, GV Instruments, Wythenshaw, Manchester, UK and $^{18}$O, AP2003, Analytical Precision Ltd, Northwich, Cheshire, UK). Laboratory reference standards which have previously been calibrated against the Vienna-Standard Mean Ocean Water and Vienna-Standard Light Antarctic Precipitate international standards, were measured alongside all samples.

Resting metabolic rate was measured by respired gas analysis during a fifteen minute rest test towards the start of both clinic visits (OxyconPro, Jaeger), which took place in the afternoon. A five-breath running median was calculated through the breath-by-breath data, and the lowest observed average metabolic rate over a consecutive five

minute window was used in analyses. This was scaled down by 6% to compensate for afternoon elevation of resting metabolic rates [36]. Basal metabolic rate was estimated by three established equations [40, 60, 87], which differ in their utilisation of body composition information, and the measured resting metabolic rate closest to the average estimated value was selected. Their final 24-hour resting metabolic rate was intergrated over awake hours and further scaled down by 5% proportional to their mean sleep duration, which was derived from their sleep diary [3].

A food frequency questionnaire was administered at the second clinic visit, which estimates dietary intake over the past year [9], and was processed using FETA [59]. Total energy intake was normalised by the total energy expenditure from doubly labelled water, to approximate the diet induced thermogenesis fraction. The calorie-weighted macronutrient profile was used to approximate the food quotient, which was used as a proxy for the respiratory quotient in the calculation of total energy expenditure [14].

### 4.2.2 Data pre-processing

The raw acceleration signals collected at the wrists and thigh were downsampled from their original 100 Hertz to 25 Hertz by linear interpolation, which reduced the dimensionality of the data whilst preserving the frequency range of possible human movement. Data windows containing non-wear from either device were excluded; non-wear for the raw accelerometer devices was defined as the device being motionless (standard deviation of acceleration in all axes $<$ 10 milli-g) continuously for an hour or longer, and non-wear for the combined sensor defined as consecutive zero movement for 90 minutes or longer and simultaneous with a non-physiologically plausible heart rate.

The individually-calibrated activity energy expenditure signal was upsampled from its original fifteen-second resolution to one-second resolution by linear interpolation. The two signals were then merged together based on the real timestamps recorded sepa-

rately by both devices.

### 4.2.3 Deep neural network models

The learning task was formulated as a vector-to-vector prediction; the input to each of the models was a two-dimensional vector containing fifteen seconds of triaxial acceleration data (3x375) (three axes, and 15 seconds x 25 samples = 375), and the output was a one-dimensional vector of second-level energy expenditure (1x15). For clarity, a diagram illustrating this arrangement is given in Figure 4.2.

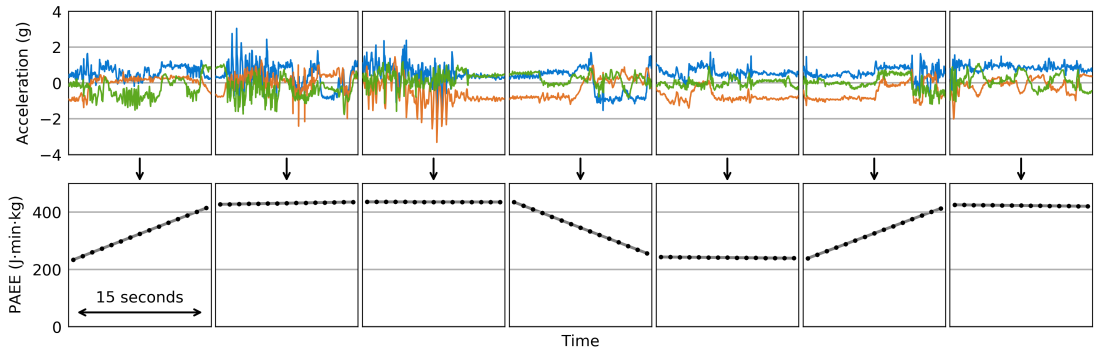


Figure 4.2: A visual illustration of the learning task. The input to the neural network is the raw acceleration signal, chopped into non-overlapping fifteen second windows. The output is the contemporaneous activity energy expenditure signal in the same time window, after linear interpolation to one-second resolution.

Similar to previous examples of neural networks used in human activity recognition [62, 35, 33, 65], the neural network topology used in these analyses consisted of convolutional layers to extract features from the input signal, long short-term memory layers to model time dependencies in those features, and fully-connected layers to ultimately interpret those features. Specifically, there were 5 convolutional layers with 128 filters per layer, each of which was followed by a max-pooling layer and batch normalization. There was one long short-term memory layer with 2048 units, followed by a batch normalization layer. Finally, there were 3 fully-connected layers with 4096 neurons per

layer, each of which was followed by a batch normalization layer. A diagram illustrating the model is given in Figure 4.3, and code to produce the model is given in the Supplementary material.

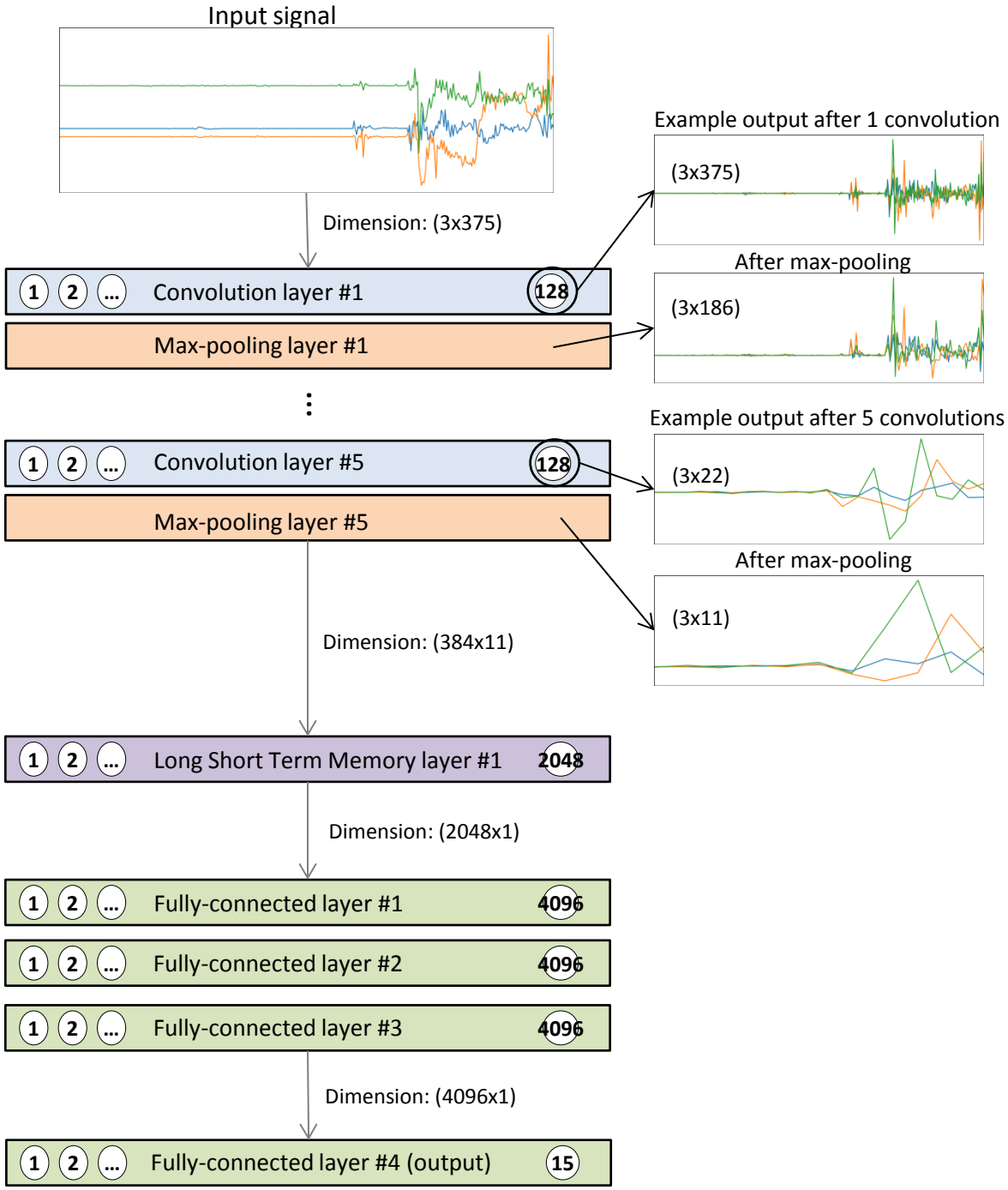


Figure 4.3: Schematic of the neural network. The convolution layers extract features, and the pooling layers collapse those along the time axis. Each unit in the LSTM layer models a response to those features. The fully-connected layers perform the final inferences; the final layer outputs 1 number for each second of input data.

### 4.2.4 Model derivation

One site-specific neural network was trained for each of the measured anatomical sites; the dominant wrist, non-dominant wrist and thigh. An identical neural network architecture was used for each experiment, as described previously. The networks were all trained in a minibatch style of 64 samples per minibatch, using the Adam optimizer [47] with a learning rate of 5x10$^{-5}$, which was always halved at the end of each epoch (an epoch in deep learning terminology is defined as a full pass over one random permutation of the entire training dataset, and has no relation to the term often used in physical activity epidemiology). The learning objective was to minimise the mean squared error of the estimation across the output window. The models were defined and trained using Keras [20] and a TensorFlow backend [1].

Within each experiment, the participants were split into "training", "validation" or "test" groups. The "test" set consisted of all participants with at least three days of valid sensor data and a doubly labelled water measurement. The remaining participants were randomly assigned to "training" or "validation" groups, resulting in approximately equally sized training and validation datasets, and a much larger test dataset.

A snapshot of the neural network weights was saved after every 10% progress into the training data, and the training loop was always terminated after 5 epochs. This process produced 50 candidate models for each anatomical site, only one of which could be taken forward to the final evaluation against the gold-standard. They were evaluated in the holdout "validation" data sample, and assessed according to agreement with the silver-standard combined sensing estimates. The best performing model was defined as the model with the lowest Root Mean Squared Error (RMSE) and a non-significant mean estimation bias at the population level. This final model was selected for evaluation in the independent "test" dataset, as assessed with the gold-standard criterion measure.

### 4.2.5 Statistical evaluation

Each deep neural network was applied in a piecewise fashion to every non-overlapping valid window of acceleration data in the independent test dataset. The resulting second level intensity time-series was then summarised to an average-per-participant estimate by means of a diurnal adjustment technique [15], which calculates a daily average by modelling the signal as a function of time-of-day, making it robust to differences between people that arise based on wear time distribution across the day. Diurnal adjustment was also applied to the individually-calibrated combined sensing signals to facilitate a comparison at the participant level.

Additionally, combinatorial estimates were created by taking the mean average of each possible combination of the summary level estimates. This effectively created four additional estimates per person (dominant wrist and thigh, non-dominant wrist and thigh, both wrists, and both wrists and thigh) which were evaluated in an identical manner to the single-sensor estimates.

Estimates of total energy expenditure (expressed in absolute terms, MJ·day$^{-1}$) were created by adding together estimated activity energy expenditure and estimated resting energy expenditure from the simplest model (using only age, sex, height and weight) [40], and dividing the sum by 0.9 to include diet-induced thermogenesis (thereby making the assumption that diet-induced thermogenesis accounts for a fixed 10% of total energy expenditure [88]).

The estimates of both activity energy expenditure and total energy expenditure were evaluated by agreement with those derived from the doubly-labelled water measurement. An additional sensitivity analysis was performed which excluded left-handed participants. Agreement was formally tested by mean bias, 95% limits of agreement, Pearson correlations, and RMSE of the estimation. All statistical tests were performed in Python 3.6.

## 4.3 Results

A summary of the participant characteristics is given in Table 4.1, and a summary of the datasets used to train and validate the models is given in Table 4.2. The separate training datasets for the dominant wrist, non-dominant wrist and thigh consisted of approximately two million observations each. In the test group, the doubly labelled water measurements showed mean (standard deviation) total energy expenditure was 11.6 (2.3) MJ·day$^{-1}$. Resting energy expenditure was 6.6 (1.2) MJ·day$^{-1}$, and activity energy expenditure was 49.8 (16.3) kJ·day$^{-1}$·kg$^{-1}$.

| | DLW (n=100) | | | | Non-DLW (n=93) | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Min | Max | Mean | SD | Min | Max |
| Sex (% women) | 50% | | | | 41% | | | |
| Age (years) | 54.4 | 7.2 | 40 | 65 | 54 | 6.7 | 41 | 66 |
| Height (m) | 1.7 | 0.1 | 1.5 | 1.9 | 1.7 | 0.1 | 1.5 | 2.0 |
| Weight (kg) | 78.2 | 13.6 | 48.7 | 110.8 | 77.1 | 12.4 | 56.4 | 112.3 |
| BMI (kg·m$^{-2}$) | 26.5 | 3.4 | 20.4 | 36.6 | 25.9 | 2.9 | 20.4 | 35.3 |
| TEE (MJ·day$^{-1}$) | 11.6 | 2.3 | 6.5 | 16.4 | - | - | - | - |
| REE (MJ·day$^{-1}$) | 6.6 | 1.2 | 3.7 | 9.8 | - | - | - | - |
| DIT fraction | 0.10 | 0.00 | 0.08 | 0.11 | - | - | - | - |
| AEE (MJ·day$^{-1}$) | 3.9 | 1.4 | 0.7 | 7.6 | - | - | - | - |
| AEE (kJ·day$^{-1}$·kg$^{-1}$) | 49.8 | 16.3 | 8.5 | 92.6 | - | - | - | - |
| $k_O$ (vSMOW·day$^{-1}$) | 0.119 | 0.03 | 0.066 | 0.257 | - | - | - | - |
| $k_H$ (vSMOW·day$^{-1}$) | 0.093 | 0.028 | 0.044 | 0.228 | - | - | - | - |
| $N_O$ (moles) | 2124 | 434 | 1215 | 3131 | - | - | - | - |
| $N_H$ (moles) | 2188 | 447 | 1251 | 3224 | - | - | - | - |

Key: BMI=Body Mass Index, TEE=Total Energy Expenditure, REE=Resting Energy Expenditure, DIT=Diet-induced Thermogenesis, AEE=Activity Energy Expenditure, DW=Dominant Wrist, NDW=Non-Dominant Wrist $k_O$ and $k_H$=disappearance rates of Oxygen and Hydrogen according to DLW, respectively. $N_O$ and $N_H$=biological pool sizes of Oxygen and Hydrogen, respectively. vSMOW=Vienna-Standard Mean Ocean Water.

Table 4.1: Participant characteristics, provided separately by doubly labelled water status.

| | Train | | Validation | |
|---|---|---|---|---|
| | Participants | Observations | Participants | Observations |
| Dominant wrist | 45 | 2,036,334 | 44 | 1,979,974 |
| Non-dominant wrist | 44 | 1,950,825 | 43 | 2,026,423 |
| Thigh | 43 | 1,927,812 | 43 | 1,944,092 |

Table 4.2: Quantity of data used to train and validate the neural network models.

In Table 4.3, we provide an overview of the single best performing model from each

of the three separate training runs, according to the comparison against the silver-standard combined sensing validation dataset. A model with no significant mean bias at the population level was found for each sensor placement, and average estimation errors ranged from 16.8 to 18.3 kJ·day$^{-1}$·kg$^{-1}$. Charts illustrating the training progress over time are given in Supplementary Figures 4.7, 4.8, and 4.9; it can be seen that model performance was always highly erratic in the early stages of training for each data source, but as the learning rate was steadily lowered, performance appeared to reach a stable equilibrium. Training sessions lasted approximately 30 minutes per epoch on machines with Intel Xeon E5-2686 v4 (Broadwell) processors and single Tesla V100 graphics cards.

| | N | Bias | 95% LoA | | r | RMSE | % error |
|---|---|---|---|---|---|---|---|
| Dominant wrist | 43 | -3.6 | -37.4 | 30.2 | 0.520 | 17.4 | 29.5 |
| Non-dominant wrist | 43 | -0.16 | -36.4 | 36.1 | 0.445 | 18.3 | 32.8 |
| Thigh | 41 | 1.2 | -32.1 | 34.5 | 0.375 | 16.8 | 32.7 |

Table 4.3: Agreement between the summary-level neural network based estimates of activity energy expenditure, and the silver-standard derived from individually-calibrated combined sensing.

Tables 4.4 and 4.5 show the final performance of the models identified in Table 4.3, when taken forward and evaluated in the test dataset against gold-standard doubly labelled water, as well as the performance of combinatorial models created by averaging those estimates. Every estimate achieved a non-significant mean bias at the population level, with the exception of estimates from the dominant wrist (which overestimated by 4.2 kJ·day$^{-1}$·kg$^{-1}$, p = 0.035), and shared very similar 95% limits of agreement between approximately -19 kJ·day$^{-1}$·kg$^{-1}$ to 25 kJ·day$^{-1}$·kg$^{-1}$. Correlations and RMSEs were very similar for every model, ranging from 0.69 to 0.73 and 12.3 kJ·day$^{-1}$·kg$^{-1}$ to 11.1 kJ·day$^{-1}$·kg$^{-1}$, respectively. Comparing the performance of each model between Tables 4.4 and 4.3, it can be seen that agreement was significantly stronger with the

gold-standard measure. The results of the sensitivity analysis using only right-handed participants is given in Supplementary Table 4.7; performance was not substantially different from the main results.

| Placement | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|
| Activity energy expenditure [via primary REE] | | | | | | |
| Dominant wrist | 97 | 4.2* | -18.5 | 26.9 | 0.687 | 12.3 |
| Non-dominant wrist | 97 | 3.3 | -18.5 | 25.1 | 0.730 | 11.5 |
| Thigh | 92 | 2.0 | -21.2 | 25.1 | 0.683 | 11.9 |
| Both wrists | 94 | 3.5 | -18.1 | 25.1 | 0.720 | 11.5 |
| Non-dominant wrist & Thigh | 90 | 2.3 | -19.4 | 23.9 | 0.725 | 11.2 |
| Dominant wrist & Thigh | 89 | 2.6 | -19.7 | 24.9 | 0.691 | 11.6 |
| Both wrists & Thigh | 87 | 2.3 | -19.0 | 23.6 | 0.716 | 11.1 |
| Activity energy expenditure [via measured REE only] | | | | | | |
| Dominant wrist | 97 | 5.6* | -18.8 | 30.1 | 0.651 | 13.6 |
| Non-dominant wrist | 97 | 4.6* | -18.8 | 28.1 | 0.694 | 12.8 |
| Thigh | 92 | 3.3 | -22.2 | 28.7 | 0.623 | 13.3 |
| Both wrists | 94 | 4.8* | -18.2 | 27.9 | 0.688 | 12.7 |
| Non-dominant wrist & Thigh | 90 | 3.5 | -20.0 | 26.9 | 0.680 | 12.4 |
| Dominant wrist & Thigh | 89 | 3.9 | -20.5 | 28.3 | 0.639 | 13.0 |
| Both wrists & Thigh | 87 | 3.5 | -19.4 | 26.5 | 0.676 | 12.2 |

Table 4.4: Agreement between the summary-level neural network based estimates of activity energy expenditure, and the gold-standard derived from doubly labelled water. An asterisk (*) next to a bias value indicates statistical significance according to a paired t-test ($p < 0.05$).

| Placement | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|
| Total energy expenditure | | | | | | |
| Dominant wrist | 97 | 0.423 | -1.6 | 2.4 | 0.903 | 1.1 |
| Non-dominant wrist | 97 | 0.353 | -1.6 | 2.3 | 0.909 | 1.1 |
| Thigh | 92 | 0.229 | -1.8 | 2.3 | 0.898 | 1.1 |
| Both wrists | 94 | 0.366 | -1.6 | 2.3 | 0.909 | 1.1 |
| Non-dominant wrist & Thigh | 90 | 0.259 | -1.7 | 2.2 | 0.910 | 1.0 |
| Dominant wrist & Thigh | 89 | 0.276 | -1.7 | 2.2 | 0.906 | 1.0 |
| Both wrists & Thigh | 87 | 0.262 | -1.7 | 2.2 | 0.912 | 1.0 |

Table 4.5: Agreement between the summary-level neural network based estimates of total energy expenditure, and the gold-standard derived from doubly labelled water. An asterisk (*) next to a bias value indicates statistical significance according to a paired t-test ($p < 0.05$).

In Figures 4.4 and 4.5 we use Bland-Altman plots to show the agreement between the summary-level estimates of activity energy expenditure and total energy expenditure from each of the final models and the gold-standard observations derived from doubly labelled water. The sloping trend in Figure 4.5 suggests the models overestimate in the least active and underestimate in the most active, but there appears to be no such trend in total energy expenditure estimates.

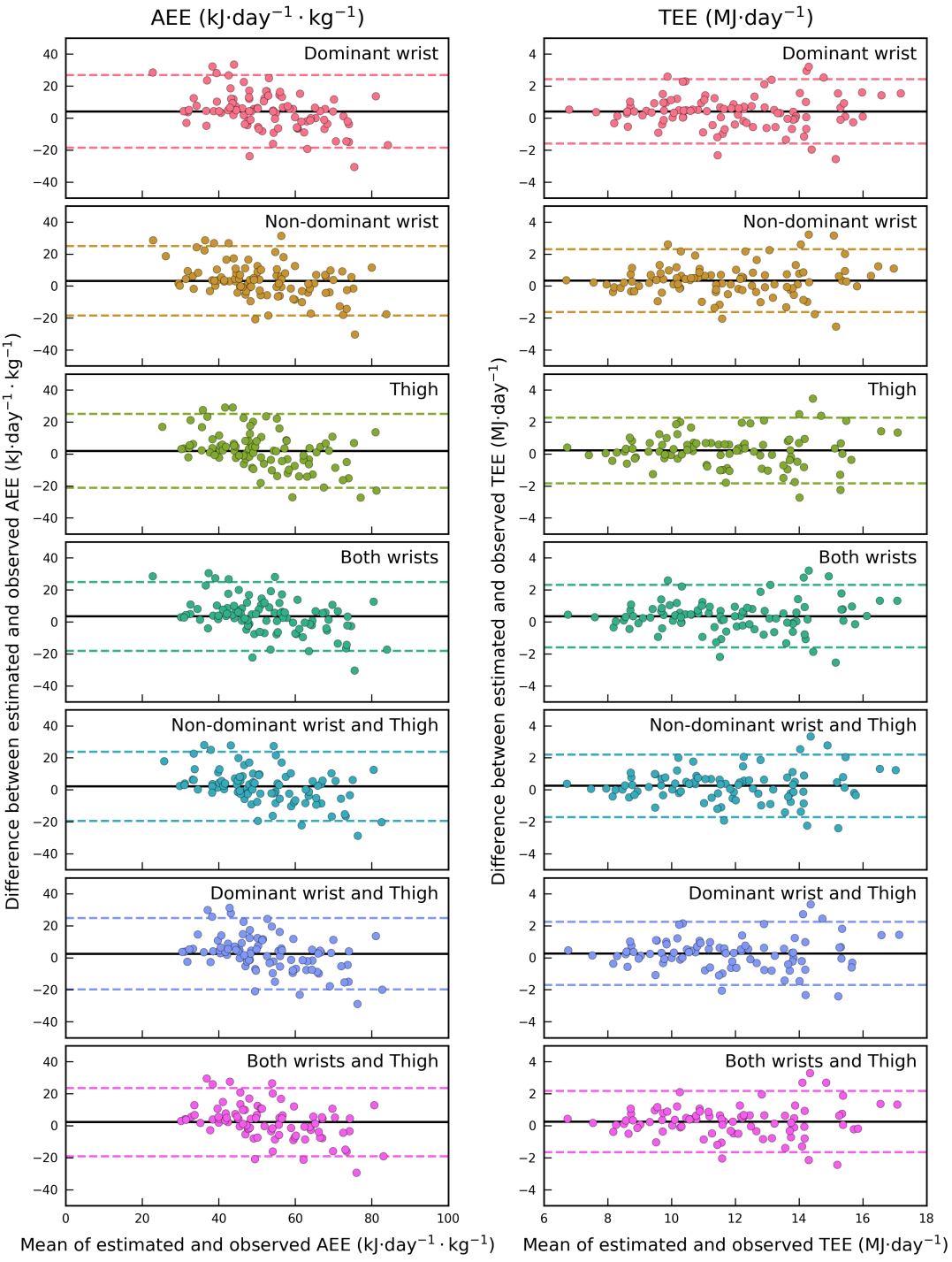


Figure 4.4: Bland-Altman plots showing the agreement between each of the summary-level estimates of activity energy expenditure with gold-standard observations derived from doubly labelled water, where the X-axis indicates the mean of measured and observed values.

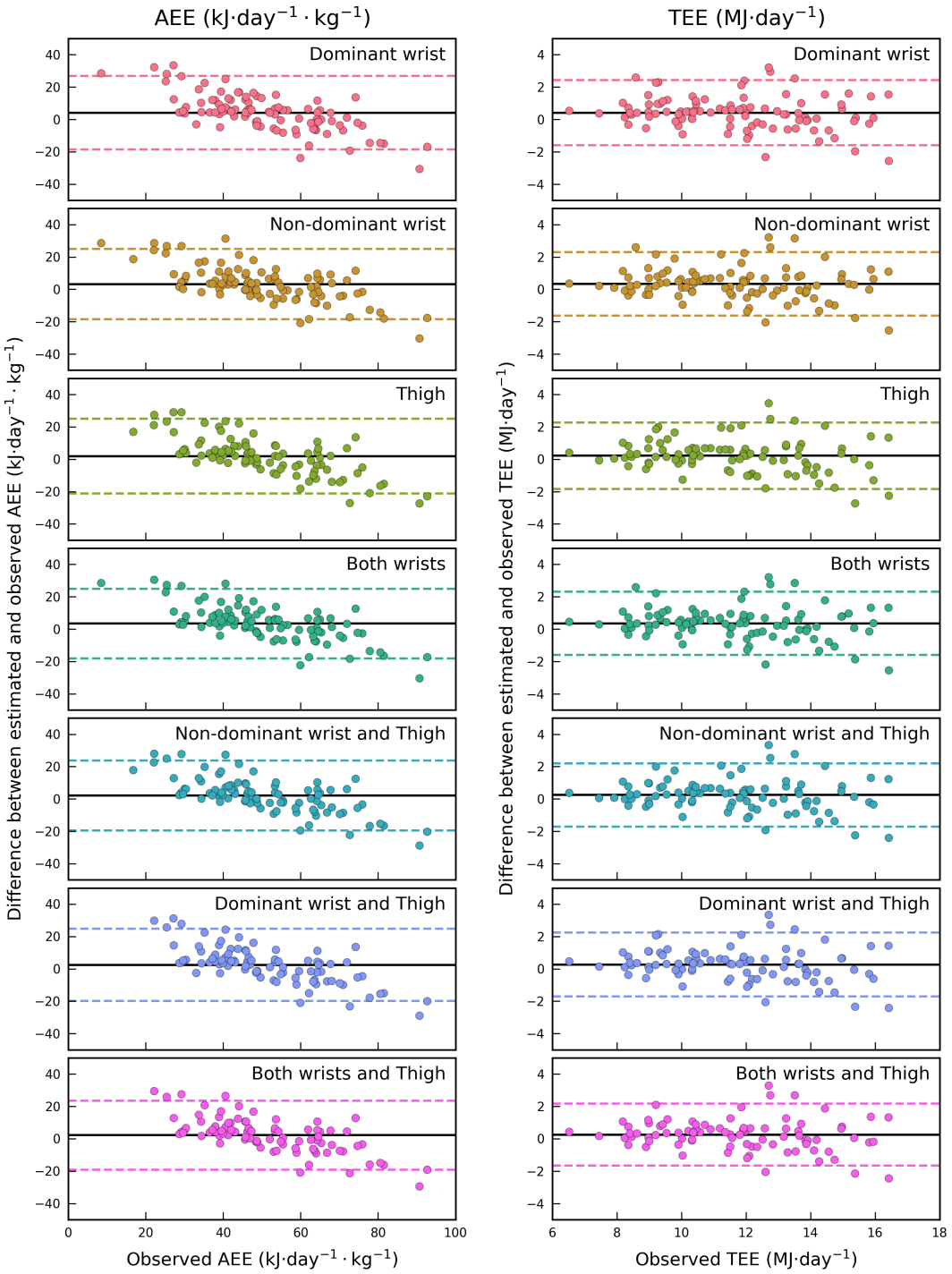


Figure 4.5: Bland-Altman plots showing the agreement between each of the summary-level estimates of activity energy expenditure with gold-standard observations derived from doubly labelled water, where the X-axis indicates the gold-standard observed value.

Table 4.6 shows the results of ensembling the summary-level estimates of activity energy expenditure from both the newly-derived neural networks and their original movement intensity counterparts established previously [91]. The slight tendency towards underestimation of the movement intensity models is complemented by an opposite tendency towards overestimation in the neural network models, resulting in population-level biases even closer to zero and an overall reduction in RMSE. For reference, in Figure 4.6 we illustrate all of the pairwise correlations between the neural network estimates of activity energy expenditure, and the original movement intensity estimates. For the wrist-based models, correlations between the neural network estimates and the movement intensity based estimates were very high (r=0.95 and 0.94 for the dominant wrist and non-dominant wrist estimates, respectively). The thigh-based estimates were the least correlated with the rest, but comparing the two thigh models, the neural network estimates were much more strongly correlated with the wrist-based estimates than the original movement based estimates.

| Placement | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|
| Activity energy expenditure [via primary REE] | | | | | | |
| Dominant wrist | 97 | 1.2 | -22.0 | 24.3 | 0.673 | 11.8 |
| Non-dominant wrist | 97 | 0.9 | -21.4 | 23.2 | 0.713 | 11.4 |
| Thigh | 91 | -1.1 | -24.9 | 22.8 | 0.660 | 12.1 |
| Both wrists | 94 | 0.8 | -21.4 | 23.0 | 0.701 | 11.3 |
| Non-dominant wrist & Thigh | 89 | -0.5 | -22.5 | 21.6 | 0.715 | 11.2 |
| Dominant wrist & Thigh | 88 | -0.4 | -23.2 | 22.4 | 0.675 | 11.6 |
| Both wrists & Thigh | 86 | -0.5 | -22.2 | 21.3 | 0.703 | 11.1 |
| Activity energy expenditure [via measured REE only] | | | | | | |
| Dominant wrist | 97 | 2.6 | -22.2 | 27.4 | 0.639 | 12.9 |
| Non-dominant wrist | 97 | 2.2 | -21.6 | 26.1 | 0.681 | 12.3 |
| Thigh | 91 | 0.2 | -25.6 | 25.9 | 0.613 | 13.1 |
| Both wrists | 94 | 2.1 | -21.4 | 25.7 | 0.672 | 12.1 |
| Non-dominant wrist & Thigh | 89 | 0.7 | -22.9 | 24.2 | 0.678 | 12.0 |
| Dominant wrist & Thigh | 88 | 0.8 | -23.9 | 25.5 | 0.631 | 12.6 |
| Both wrists & Thigh | 86 | 0.7 | -22.5 | 23.9 | 0.669 | 11.8 |

Table 4.6: Agreement between the summary-level neural network based estimates of activity energy expenditure, and the gold-standard derived from doubly labelled water. An asterisk (*) next to a bias value indicates statistical significance according to a paired t-test ($p < 0.05$).

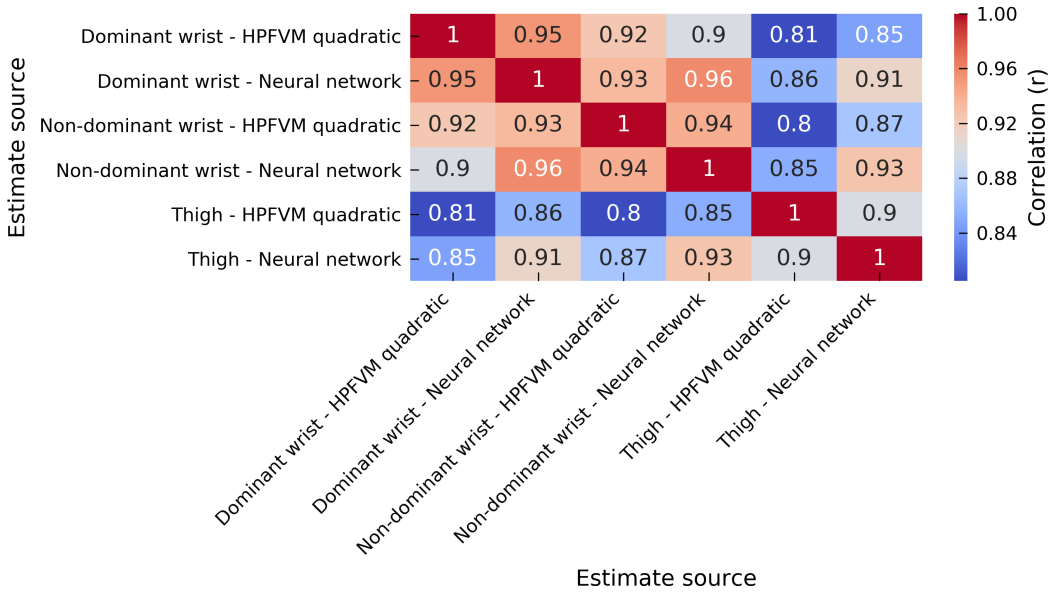


Figure 4.6: The pairwise correlations between the summary-level estimates of activity energy expenditure, according to the newly derived neural network models and the previously established movement intensity models.

## 4.4 Discussion

In this work we have introduced a novel methodology for inferring the activity energy expenditure of free-living adults measured by wrist and thigh accelerometry, based on deep convolutional neural networks. We have evaluated these deep neural network based approaches in an independent sample by agreement with doubly labelled water, and observed small and mostly non-significant mean biases at the population level, low average estimation errors (11.6 kJ·day$^{-1}$·kg$^{-1}$) and high correlations with the criterion (r = 0.7). In combination with predicted resting energy expenditure [40], estimates of total energy expenditure also agreed strongly with the criterion (r = 0.9, RMSE = 1 MJ·day$^{-1}$). These performances are an improvement upon the current movement

intensity based approaches [64, 70].

Using this same dataset, we have previously reported on the absolute validity of estimating activity energy expenditure from wearable sensors by interpreting movement intensity [91, 89]. Every neural network in this work was an improvement upon its movement intensity counterpart; correlations were on average 0.05 higher (0.71 vs 0.66), and RMSEs were on average 0.8 kJ·day$^{-1}$·kg$^{-1}$ lower (11.6 vs 12.4 kJ·day$^{-1}$·kg$^{-1}$), with tighter limits of agreement in all cases. The greatest improvements were found for thigh acceleration, where correlation with the reference improved from 0.60 to 0.68, and RMSE lowered from 13.6 to 11.9 kJ·day$^{-1}$·kg$^{-1}$ (a 12% reduction). These improvements to estimating activity energy expenditure did not translate to noticeably improved estimates of total energy expenditure; correlations were around r=0.9 and RMSEs were approximately 1 MJ·day$^{-1}$, which was very similar to those from the movement intensity estimates.

For each of the newly derived deep neural networks, we observed a markedly better performance in the test set over the validation set, which would not ordinarily be expected. We suspect this is attributable to the known higher degree of normally-distributed random measurement error in the validation data [14], as it was acquired by the silver-standard combined sensing methodology. This would naturally result in an attenuation of the agreement between the estimates, without compromising the central tendency of the estimate.

Previously, hybrid machine learning approaches have been used to estimate activity energy expenditure from triaxial acceleration data measured at the hip, where the acceleration was expressed in the proprietary units determined by the manufacturer [52]. These models were trained using data collected in 6 participants, where a researcher observed their activities (sitting, standing, walking, etc) during free-living for up to 30 hours in 10-hour long sessions. A large recent evaluation of this approach, assessed

by agreement with doubly labelled water in 683 participants, found a significant mean estimation bias of -8.8 kJ·day$^{-1}$·kg$^{-1}$ and 95% limits of agreement between -38 and 20 kJ·day$^{-1}$·kg$^{-1}$, a 22% underestimation [55]. The relatively poor performance compared to the present results may be attributable to their training dataset, which was much smaller and less precisely labelled. Performance may also be affected by sensor location, and the on-board processing of the data by the device into proprietary units.

A common criticism of deep neural networks is that they yield highly complex models which are virtually inscrutable, which impedes us from learning anything meaningful by examining their many millions of coefficients. This model inscrutability means we cannot guarantee the model will perform sensibly and reliably when presented with unfamiliar data, and the greater complexity of these models also increases their potential for population specificity. It is important that future work examines their performance in a broader range of individuals. If specific populations are identified wherein the model performance is sub-optimal, one potential remedy is to acquire relevant training data to resume the learning process and further refine the models. In machine learning terminology, this is known as fine-tuning, and it is used when only a limited quantity of data is available for a specific problem, but a model has already been trained to solve a similar problem where data is more readily available.

By contemporary deep learning standards, the network topologies used herein would likely not be considered “deep”; in the imaging domain, models with hundreds of layers are now commonplace [37]. New architectures and training methodologies are regularly proposed which further improve performance. As such, these are most likely not the definitive neural networks to estimate activity energy expenditure; they are merely a starting point and a strong benchmark against which to evaluate future models. Training and even simply applying these models to raw acceleration data is a computationally expensive process, which is made practical only by the massive parallelisation available

with modern graphical processing units. We recommend that anyone intending to utilise our models does so using an accelerated computing environment.

In conclusion, we have trained deep neural networks to predict activity energy expenditure from raw acceleration data measured at three different anatomical sites during free-living, and those estimates when integrated over time agreed strongly with measurements by a gold-standard criterion, doubly labelled water. Our results suggest that these models provide the most accurate and precise estimates of free-living activity energy expenditure from accelerometry to date [64, 70]. A large number of epidemiological studies have already collected raw acceleration data in free-living conditions [25, 78, 22], and applying this inference scheme will further enhance their utility by providing a more precise characterisation of activity energy expenditure profiles.

## 4.5 Supplementary material

### 4.5.1 Code to create model

```
from keras.models import Sequential
from keras.layers import LSTM, Dense, Reshape, Permute, Flatten, Conv2D,
                                    MaxPooling2D, BatchNormalization
from keras import optimizers
from keras import regularizers

# Properties of Conv2D layers
conv = dict(kernel_size=(1, 3), padding="same",
  activation="relu", kernel_regularizer=regularizers.l2(0.01))

num_conv_layers = 5
```

```
num_lstm_layers = 1
num_dense_layers = 3
n_conv = 128
n_lstm = 2048
n_dense = 4096

model = Sequential()

model.add(BatchNormalization(input_shape=(3, 375, 1)))

# Convolution layers
for i in range(num_conv_layers):
  model.add(Conv2D(n_conv, **conv))
  model.add(MaxPooling2D(pool_size=(1, 2)))
  model.add(BatchNormalization())

# Put the time dimension in the right place
model.add(Permute((2, 1, 3)))
model.add(Reshape((-1, (n_conv*3))))

# LSTM layers
for i in range(num_lstm_layers):
  rs = (i < num_lstm_layers-1)
  model.add(LSTM(n_lstm, return_sequences=rs,
    activation="relu", kernel_regularizer=regularizers.l2(0.01)))
  model.add(BatchNormalization())

# Dense layers
for i in range(num_dense_layers):
  model.add(Dense(n_dense, activation="relu",
    kernel_regularizer=regularizers.l2(0.01)))
  model.add(BatchNormalization())
```

```
# Final output
model.add(Dense(15, activation="relu",
  kernel_regularizer=regularizers.l2(0.01)))

optimizer = optimizers.Adam(lr=0.00005)
model.compile(loss='mean_squared_error', optimizer=optimizer)

model.summary()
```

### 4.5.2 Instructions to participants: Monitor placement

One monitor should be positioned on each wrist. The monitors should be positioned so that the engraving on the band is on the left hand side (as pictured below). The monitors have been coloured to indicate which monitor should be worn on which wrist. Please wear the black band on your RIGHT wrist and the green band on your LEFT wrist.

Both wrist monitors should be worn just above the joint so that when the joint is flexed, the monitors neither inhibit joint movement nor are uncomfortable. Both monitors should

remain in the position in which they were placed in the clinic, and should retain a snug fit and not allowed to rotate.

You have been provided with a thigh worn monitor. Please try and wear the monitor continuously for 8 days and nights. During this time, please carry on with all your activities as usual.

Placement on the Thigh: The monitor should be positioned on the right thigh, if for any reason you need to remove the monitor, please replace it as shown at your clinic visit and note down on the physical activity monitor diary sheet when it was taken off and put back on.

### 4.5.3 Sensitivity analysis

| Placement | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|
| Activity energy expenditure [via primary REE] | | | | | | |
| Dominant wrist | 90 | 4.4* | -18.2 | 27.0 | 0.706 | 12.3 |
| Non-dominant wrist | 90 | 3.5 | -18.3 | 25.3 | 0.742 | 11.6 |
| Thigh | 86 | 2.1 | -21.1 | 25.4 | 0.690 | 12.0 |
| Both wrists | 87 | 3.7 | -17.8 | 25.2 | 0.736 | 11.5 |
| Non-dominant wrist & Thigh | 84 | 2.5 | -19.2 | 24.2 | 0.734 | 11.3 |
| Dominant wrist & Thigh | 83 | 2.8 | -19.5 | 25.1 | 0.702 | 11.6 |
| Both wrists & Thigh | 81 | 2.5 | -18.7 | 23.8 | 0.728 | 11.1 |

Table 4.7: Agreement between the summary-level neural network based estimates of activity energy expenditure, and the gold-standard derived from doubly labelled water, in only right-handed participants. An asterisk (*) next to a bias value indicates statistical significance according to a paired t-test ($p < 0.05$).

### 4.5.4 Validation performance

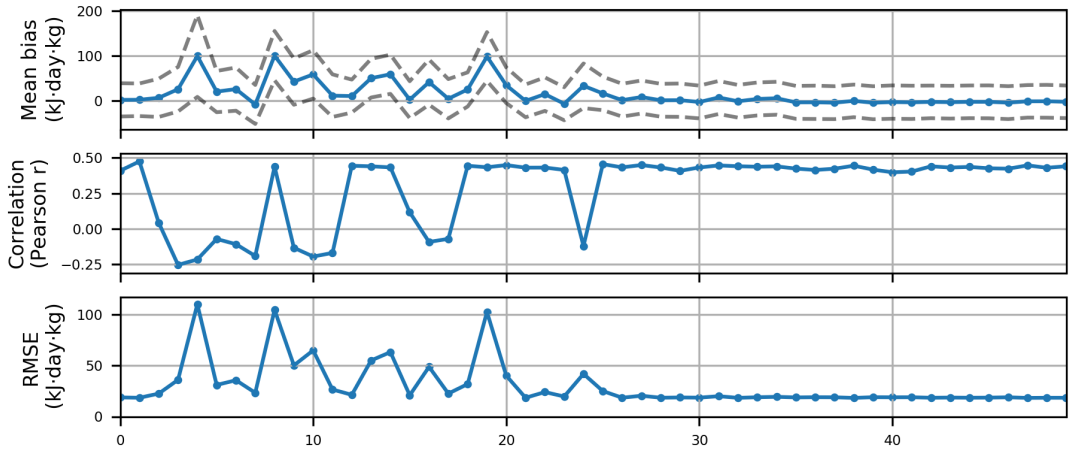


Figure 4.7: Performance of the non-dominant wrist models throughout training, evaluated by agreement with combined-sensing.

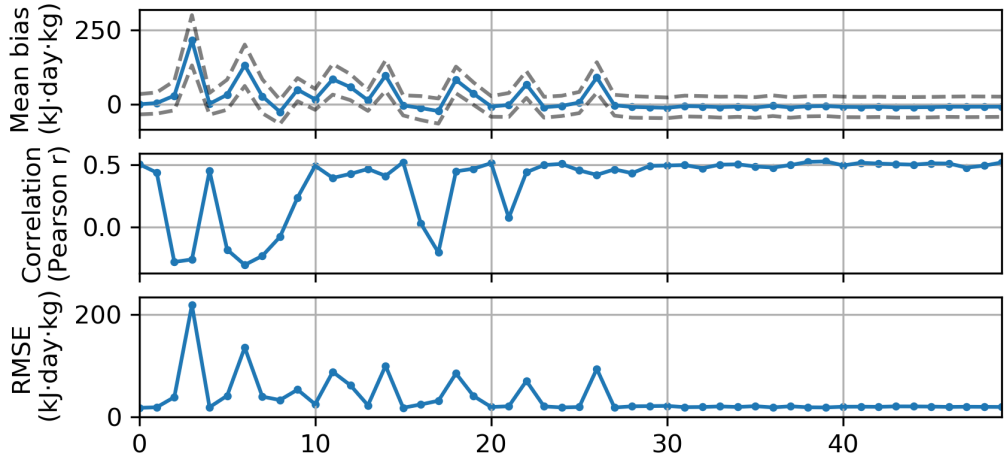


Figure 4.8: Performance of the dominant wrist models throughout training, evaluated by agreement with combined-sensing.

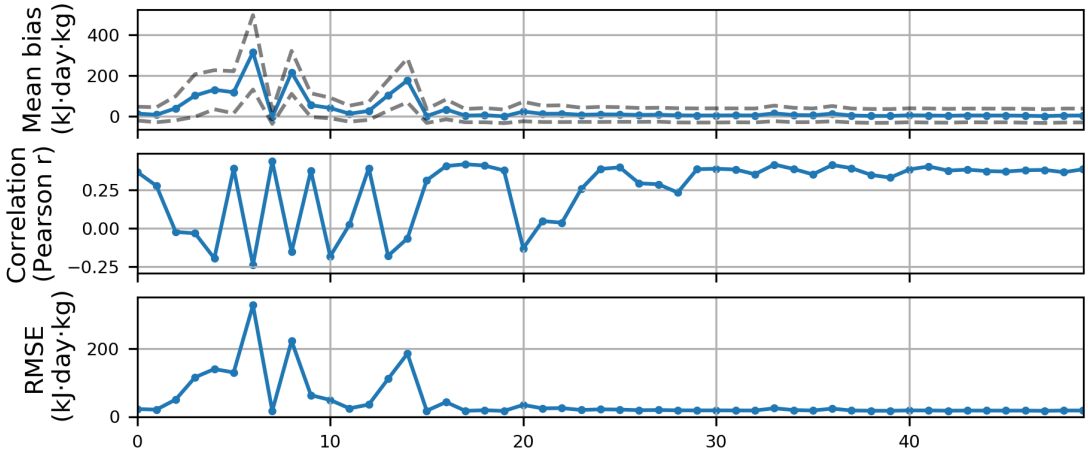


Figure 4.9: Performance of the thigh models throughout training, evaluated by agreement with combined-sensing.

Chapter 5

# Assessment of activity energy expenditure by wrist accelerometry in sub-Saharan Africa: The Cameroon study

*Introduction: Wrist acceleration has been employed in a diverse range of studies across the world to capture free-living physical activity. Models to estimate activity energy expenditure from such data have been developed and evaluated, but this methodological work has largely been conducted in the developed world, and it is unclear if these approaches are applicable in a wider setting.*

*Methods: 713 Cameroonian adults wore a triaxial accelerometer on the non-dominant wrist, simultaneously with a combined heart rate and movement sensor on the trunk, which was interpreted using a step-based exercise test to produce an individually-calibrated activity energy expenditure signal. Using data from 80 participants, a neural network was trained to estimate the activity energy expenditure signal from the wrist signal, and another network originally trained on British data was fine-tuned using the same dataset. Alongside two other models derived in British adults, these models were applied in the remaining 633 participants, and agreement was then assessed between those and the combined sensor.*

*Results: Mean activity energy expenditure from the combined sensor was estimated at 47.1 (SD=19.6) kJ·day$^{-1}$·kg$^{-1}$. The models derived in British adults, and the neural network trained from scratch, over-estimated by 11.2, 15.4, and 9.4 kJ·day$^{-1}$·kg$^{-1}$, respectively. A non-significant mean bias was found only for the fine-tuned network. However, all estimates were correlated with the criterion ($r > 0.6$), and similar associations were found with demographics such as age, sex, and BMI.*

*Discussion: The relationship between wrist acceleration intensity and activity energy expenditure is different in British and Cameroonian adults, causing models derived in British data to overestimate at the population level in Cameroonians. Further work is required before population-level comparisons of activity energy expenditure can be made from data collected exclusively by wrist accelerometry.*

## 5.1 Introduction

There is a worldwide trend towards greater prevalance of obesity and related metabolic disorders [2]. We are simultaneously witnessing a rapid transition towards urban living in the developing world, and corresponding changes in dietary and activity patterns. In order to fully understand these changes and to investigate hypothesised causal relationships [5], we need to be able to accurately and reliably measure the individual components of energy intake and energy expenditure which ultimately contribute to overall energy balance. It is especially important that our measurement instruments can be utilised at a population level in parts of the world currently undergoing this transition [30], where the most valuable insights may be found. Tools such as wrist-worn triaxial accelerometers have proven to be a feasible option to capture physical activity in free-living conditions in large-scale studies [25, 78, 22], due to the ease of deployment and high acceptability by participants [83]. However, approaches for the assessment of activity energy expenditure using wrist acceleration data have been developed and subsequently validated almost exclusively in developed Western countries, and it remains unknown if these inferences are equally applicable in other cultures and countries.

An accelerometer logs three-dimensional acceleration at a high frequency, which when attached to a participant produces a rich biomechanical signal describing human motion in extraordinary detail. Accelerometry is therefore a measurement of movement rather than energy expenditure per se, but there is a long history of inferring energy expenditure from movement signals [64]. In our previous work, we have proposed two different modelling approaches for the estimation of activity energy expenditure from wrist acceleration data collected in free-living [91] (and chapter 4). The first approach uses traditional parametric equations, which summarise the raw signal by deriving met-

rics of wrist movement intensity, and multi-level regression methods were used to derive linear and quadratic equations which relate that to activity energy expenditure intensity. The second approach uses deep neural networks to model the highly nonlinear relationship from the raw acceleration data itself to activity energy expenditure, relying on its internal transformations to derive features of explanatory power. Both of these methods have been evaluated by integrating the estimated time-series and assessing their agreement with activity energy expenditure derived from doubly labelled water [89] (and chapter 3), which is a gold-standard measure of total energy expenditure in free-living humans [72, 71].

When the deep neural network model was derived to estimate activity energy expenditure in British adults, it was speculated that there was a greater potential for it to be population specific, as the relative complexity of the model may make it more likely to fit movement patterns that are most common in the training population. It was suggested that population-specificity could be remedied by "fine-tuning": the practice of continuing model training in a small sample from the target population.

This study had two main aims. Firstly, we aimed to apply our estimation models to a dataset collected in Cameroon, and to examine their agreement with a silver-standard criterion measure of activity energy expenditure. Secondly, we aimed to derive two deep neural networks to estimate activity energy expenditure in Cameroonian adults; the first by fine-tuning our original neural network in Cameroonian data, and the second trained from scratch using the same dataset.

## 5.2 Methods

The Cameroon study was established to quantify free-living physical activity in population-based studies based in Africa [5]. This analysis was conducted using a dataet collected as part of the Cameroon 2 study, which was established to study secular trends in Cameroon. The ongoing rapid urbanisation of Cameroon was reflected in the study design; participants were recruited from within a forest region and a savannah region (Yaoundé and Bamenda, respectively), and within each region the aim was to recruit a roughly equal number of participants from designated urban and rural settings. The study received ethics approval from the Cameroon National Ethics Committee, and all participants provided written informed consent.

Wrist acceleration was captured using a triaxial accelerometer (GeneActiv, ActivInsights, Cambridge, UK) worn on the non-dominant wrist, measuring at 60 Hertz for up to 7 days with a dynamic range of $\pm$ 8 g. The participants simultaneously wore a combined heart rate and movement sensor (Actiheart, camNtech, Cambridge, UK) which measured heart rate and trunk acceleration in one-minute epochs.

Prior to the free-living measurement period, the participants performed an eight minute long step test [13], in order to establish their heart rate at standardised levels of energy expenditure. The step test results were used to interpret the subsequent free-living combined sensing signal to produce an individually-calibrated activity energy expenditure signal [12, 11, 13]. The validity of this inference scheme was investigated as a substudy in a previous wave of this study [6], wherein it was found that the estimates have no significant mean bias at the population level, with a Root Mean Squared Error of 29.3 kJ·day$^{-1}$·kg$^{-1}$.

### 5.2.1 Derivation of new models

A total sample of 713 participants had both wrist acceleration and individually-calibrated combined sensing data available. A sub-sample of 80 of those participants were randomly selected to be held back for model derivation, such that there were 20 people in each combination of men/women and urban/rural. Half of each group was randomly assigned to either be part of the “training” sample, or the independent “validation” sample used afterwards to select the best model. The remaining participants (n=633) are hereafter referred to as the ”test” set.

Two neural networks of an identical topology were trained, the details of which have been described elsewhere (chapter 4). Briefly, the learning problem was framed as a vector-to-vector regression problem; given fifteen seconds of raw triaxial acceleration data, estimate the simultaneous fifteen samples of activity energy expenditure signal. The topology was therefore a deep convolutional neural network design, using convolutional layers to extract features from the raw time-series input, which were subsequently fed to recurrent layers to model dependencies along the time axis, and fully-connected layers for the final interpretation.

The first network was trained from scratch after random weight initialisation, following precisely the same methodology as our previous work (chapter 4). The second was initiated with the learned weights of our original model trained on the British dataset.

Both networks were trained in a minibatch style of 64 samples per minibatch, using the Adam optimizer [47] to minimise the mean squared error of the estimation across the output window. The models were defined and trained using Keras [20] and a TensorFlow backend [1]. The new neural network was trained with an initial learning rate of $5 \times 10^{-5}$, whereas the fine-tuned network started with $1.25 \times 10^{-5}$. Consistent with our original methodology (chapter 4), both networks were trained for five full passes over the training data, halving the learning rate each time, and saving the state of each

model after every 10% progress into the training dataset. A single model was then selected from each training session by evaluating their estimates by agreement with combined sensing in the 40 “validation” participants, and choosing the model with the lowest RMSE at the group level.

### 5.2.2 Statistical evaluation

Using each estimated activity energy expenditure signal, an average-per-day estimate of activity energy expenditure (kJ·day$^{-1}$·kg$^{-1}$) was derived for each individual by diurnal adjustment [15], in order to compensate for between-individual differences introduced by time-of-day wear biases. All assessments were performed at person-level.

The population was stratified by sex, age quintiles, BMI categories (according to the World Health Organisation criteria), and by urban/rural status. Estimated activity energy expenditure was described across strata, separately for each of the five estimate sources, and linear regression models were used to characterise the joint associations between those aforementioned characteristics and each standardised activity energy expenditure estimate. Within each stratum, the four wrist-based estimates were assessed for agreement with the combined sensing estimates; agreement was assessed by evaluating mean bias, 95% limits of agreement, and Pearson correlations.

## 5.3 Results

An overview of the participant demographics is given in Table 5.1. After excluding those with insufficient data in either the wrist or combined sensing measurements, and the 80 participants used to derive new models, we were able to apply and evaluate our models in a total of 633 participants (327 from a rural setting, and 306 from an urban setting). There was a wide range of both age and BMI, from 18 to 84 years and 16 to 53 kg·m$^{-2}$,

respectively.

| | Mean | SD | Min | Median | Max |
|---|---|---|---|---|---|
| Age (years) | 40.5 | 13.3 | 18 | 40 | 84 |
| Sex (%) | 43% men, 57% women | | | | |
| BMI ($kg \cdot m^{-2}$) | 27.3 | 5.8 | 15.8 | 26.0 | 52.6 |
| Wrist acceleration (HPFVM, milli-g) | 52.8 | 12.6 | 20.1 | 51.3 | 104.9 |
| Wrist acceleration (ENMO, milli-g) | 34.1 | 9.5 | 12.2 | 32.8 | 74.3 |
| Trunk acceleration ($m \cdot s^{-2}$) | 0.12 | 0.05 | 0.01 | 0.11 | 0.43 |
| AEE - Combined sensing ($kJ \cdot day^{-1} \cdot kg^{-1}$) | 47.1 | 19.6 | 6.0 | 44.2 | 118.0 |
| AEE - HPFVM quadratic ($kJ \cdot day^{-1} \cdot kg^{-1}$) | 58.4 | 15.5 | 16.6 | 56.5 | 122.6 |
| AEE - original NN ($kJ \cdot day^{-1} \cdot kg^{-1}$) | 62.6 | 13.0 | 24.0 | 62.1 | 132.2 |
| AEE - fine-tuned NN ($kJ \cdot day^{-1} \cdot kg^{-1}$) | 47.0 | 9.3 | 16.5 | 46.6 | 79.1 |
| AEE - new NN ($kJ \cdot day^{-1} \cdot kg^{-1}$) | 56.5 | 9.4 | 30.7 | 56.0 | 96.1 |

Table 5.1: Summary descriptions of the participants in the Cameroon 2 study.

Table 5.2 shows the validation performance of the two newly-derived neural network models, determined by their agreement with the combined sensing estimates in an independent sample of 40 participants. Charts illustrating the training progress are shown in Supplementary Figures 5.5 and 5.6, respectively. Both models achieved a non-significant mean bias at the group level, and performance was very similar with correlations around r=0.57 and RMSEs of approximately 15.5 $kJ \cdot day^{-1} \cdot kg^{-1}$.

| | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|
| Fine-tuned neural network | -2.5 | -33.0 | 28.0 | 0.562 | 15.6 |
| New neural network | -0.5 | -31.1 | 30.2 | 0.576 | 15.5 |

Table 5.2: Validation performance of the two newly-derived neural networks, according to their agreement with estimates from combined sensing in the intermediate validation sub-sample of 40 participants.

Table 5.3 summarises the agreement between combined sensing and each estimate

from the non-dominant wrist, overall and by population strata. The two models derived in British adults were positively biased with respect to the silver-standard; the parametric model over-estimated by 11.3 kJ·day$^{-1}$·kg$^{-1}$, and the original neural network model over-estimated by 15.5 kJ·day$^{-1}$·kg$^{-1}$, though both estimates were similarly correlated with combined sensing (r=0.63 and 0.60, respectively). The newly-derived neural network also over-estimated by 9.3 kJ·day$^{-1}$·kg$^{-1}$, whereas the fine-tuned network was the only estimate to not have a statistically significant mean bias. In Figure 5.1, Bland-Altman plots show the agreement at participant-level between combined sensing estimates and each of the wrist estimates; the sloping trend that is more clearly visible in the two newly-derived models indicate a general regression-to-the-mean (a tendency towards overestimation in the least active, and underestimation in the most active).

| | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|
| HPFVM quadratic | 633 | 11.3* | -19.4 | 41.9 | 0.627 | 19.3 |
| Original neural network | 633 | 15.5* | -15.3 | 46.3 | 0.603 | 22.0 |
| Fine-tuned neural network | 633 | -0.1 | -30.7 | 30.5 | 0.625 | 15.6 |
| New neural network | 633 | 9.4* | -20.6 | 39.3 | 0.650 | 17.9 |

Table 5.3: Summary of agreement between combined sensing and the four wrist-based estimates of activity energy expenditure (kJ·day$^{-1}$·kg$^{-1}$). An asterisk (*) after the bias value indicates it was statistically significant according to a paired t-test ($p < 0.05$).

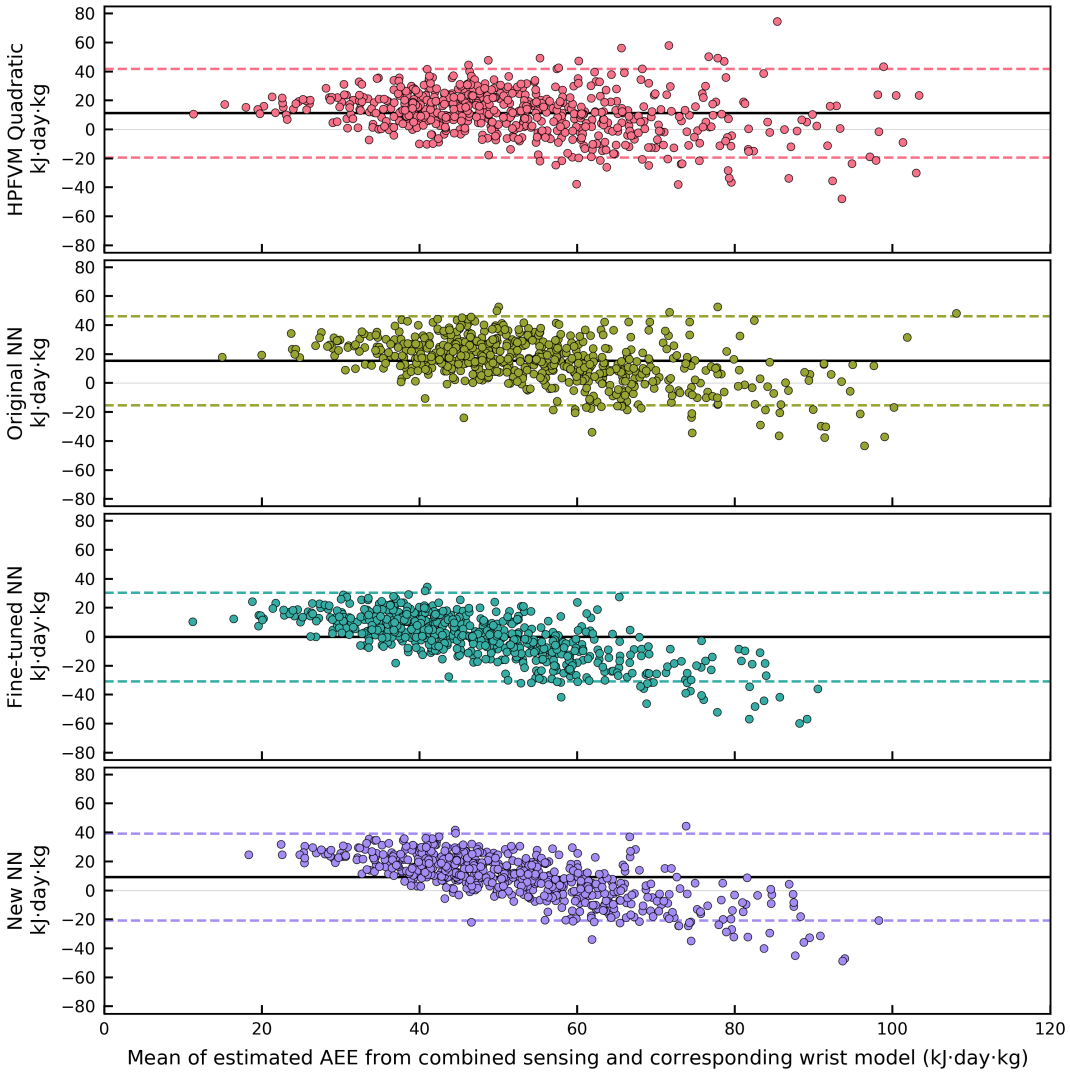


Figure 5.1: Bland-Altman plots demonstrating agreement between four wrist-based estimates of activity energy expenditure with that of combined sensing.

Agreement between combined sensing and each wrist estimate is shown in Table 5.4, shown separately for men and women and urban/rural dwellers. For brevity, agreement within the remaining population strata is illustrated in Figure 5.2. Agreement was stronger in men than in women; all wrist-based estimates were significantly more positively biased in women (on average 8 kJ·day$^{-1}$·kg$^{-1}$ higher), including the newly-

derived models from the Cameroonian training sample. Estimation bias was slightly higher on average in rural participants, but those estimates were more highly correlated with combined sensing (r=0.67 in rural versus r=0.57 in urban participants). As shown in Figure 5.2, across every estimate there were clear trends towards overestimation with older age and higher BMI.

| | N | Bias | 95% LoA | | r | RMSE |
|---|---|---|---|---|---|---|
| | | | Male participants | | | |
| HPFVM quadratic | 273 | 6.7* | -25.5 | 38.8 | 0.641 | 17.7 |
| Original NN | 273 | 10.9* | -21.7 | 43.4 | 0.616 | 19.8 |
| Fine-tuned NN | 273 | -4.9* | -36.9 | 27.2 | 0.649 | 17.0 |
| New NN | 273 | 5.3* | -26.6 | 37.2 | 0.654 | 17.1 |
| | | | Female participants | | | |
| HPFVM quadratic | 360 | 14.8* | -12.8 | 42.3 | 0.630 | 20.4 |
| Original NN | 360 | 18.9* | -8.6 | 46.5 | 0.599 | 23.6 |
| Fine-tuned NN | 360 | 3.5* | -23.9 | 30.9 | 0.610 | 14.4 |
| New NN | 360 | 12.4* | -14.5 | 39.4 | 0.634 | 18.5 |
| | | | Rural participants | | | |
| HPFVM quadratic | 327 | 12.0* | -17.2 | 41.3 | 0.677 | 19.1 |
| Original NN | 327 | 16.5* | -13.3 | 46.3 | 0.647 | 22.4 |
| Fine-tuned NN | 327 | 0.2 | -30.0 | 30.3 | 0.668 | 15.3 |
| New NN | 327 | 9.3* | -20.3 | 39.0 | 0.686 | 17.8 |
| | | | Urban participants | | | |
| HPFVM quadratic | 306 | 10.5* | -21.6 | 42.5 | 0.562 | 19.4 |
| Original NN | 306 | 14.3* | -17.3 | 46.0 | 0.547 | 21.6 |
| Fine-tuned NN | 306 | -0.4 | -31.5 | 30.8 | 0.572 | 15.9 |
| New NN | 306 | 9.4* | -21.0 | 39.8 | 0.605 | 18.1 |

Table 5.4: Agreement between combined sensing and the four wrist-based estimates of activity energy expenditure (kJ·day$^{-1}$·kg$^{-1}$), shown separately for men, women, urban and rural participants. An asterisk (*) after the bias value indicates it was statistically significant according to a paired t-test ($p < 0.05$).

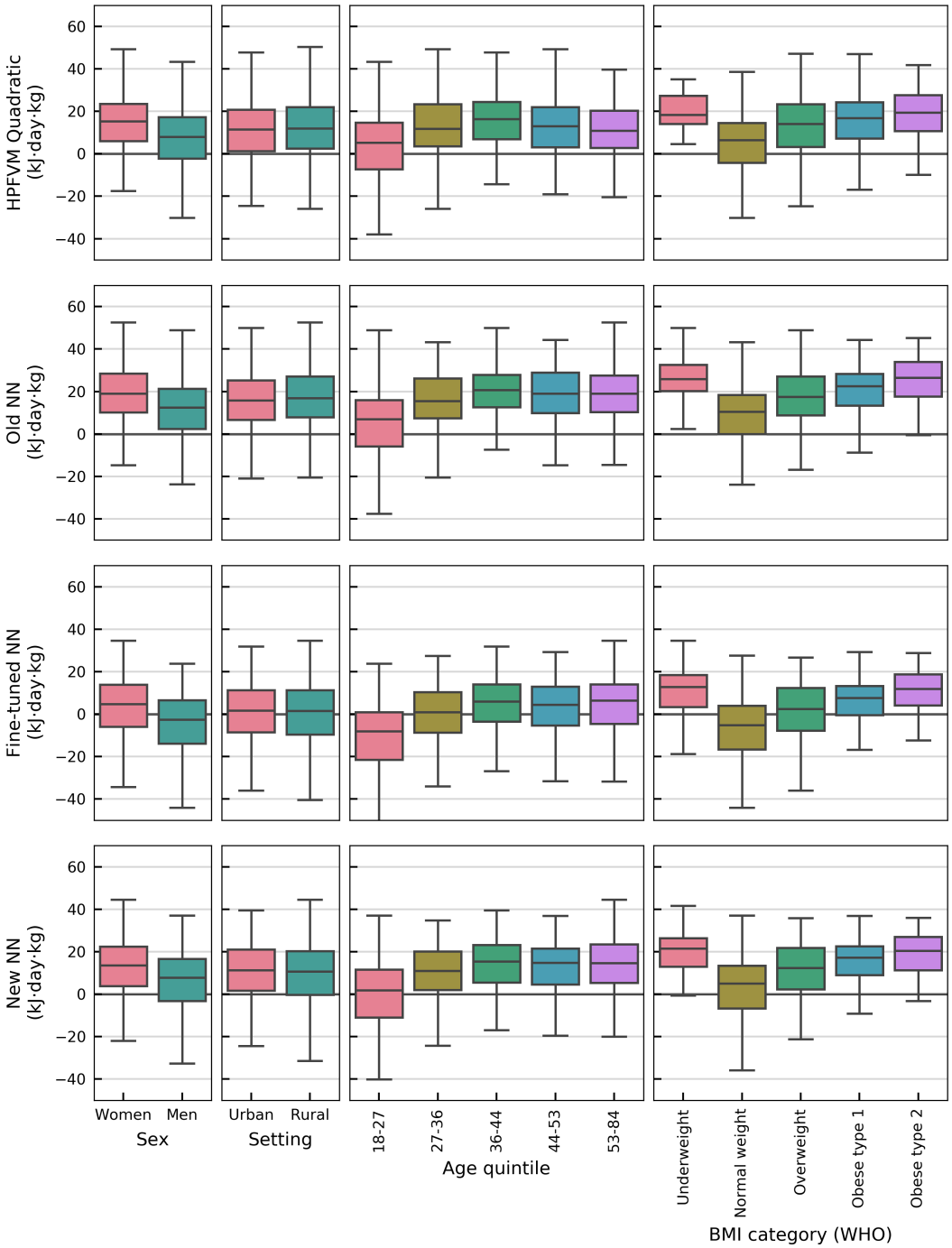


Figure 5.2: Boxplots showing the distribution of differences between activity energy expenditure estimated by combined sensing versus each of the wrist-based models, across various relevant population strata.

A series of boxplots in Figure 5.3 shows the distribution of unadjusted estimated activity energy expenditure across the various population strata, separately for each estimate source. Broadly speaking, very similar patterns appeared regardless of the estimation method; the coefficients from the regression models describing the direction and magnitude of those joint relationships are given in Table 5.5. Age and BMI were consistently negatively associated with lower activity energy expenditure across every estimate (-0.02 SD per year of age and -0.02 SD per kg·m$^2$ on average, respectively). Non-dominant wrist acceleration was not significantly different between men and women according to either ENMO or HPFVM, and of the wrist-based activity energy expenditure estimates, only the newly-derived neural network estimated significantly higher activity energy expenditure for men (0.17 SD). Trunk acceleration was higher in men (0.50 SD) but not significantly associated with urban or rural status.

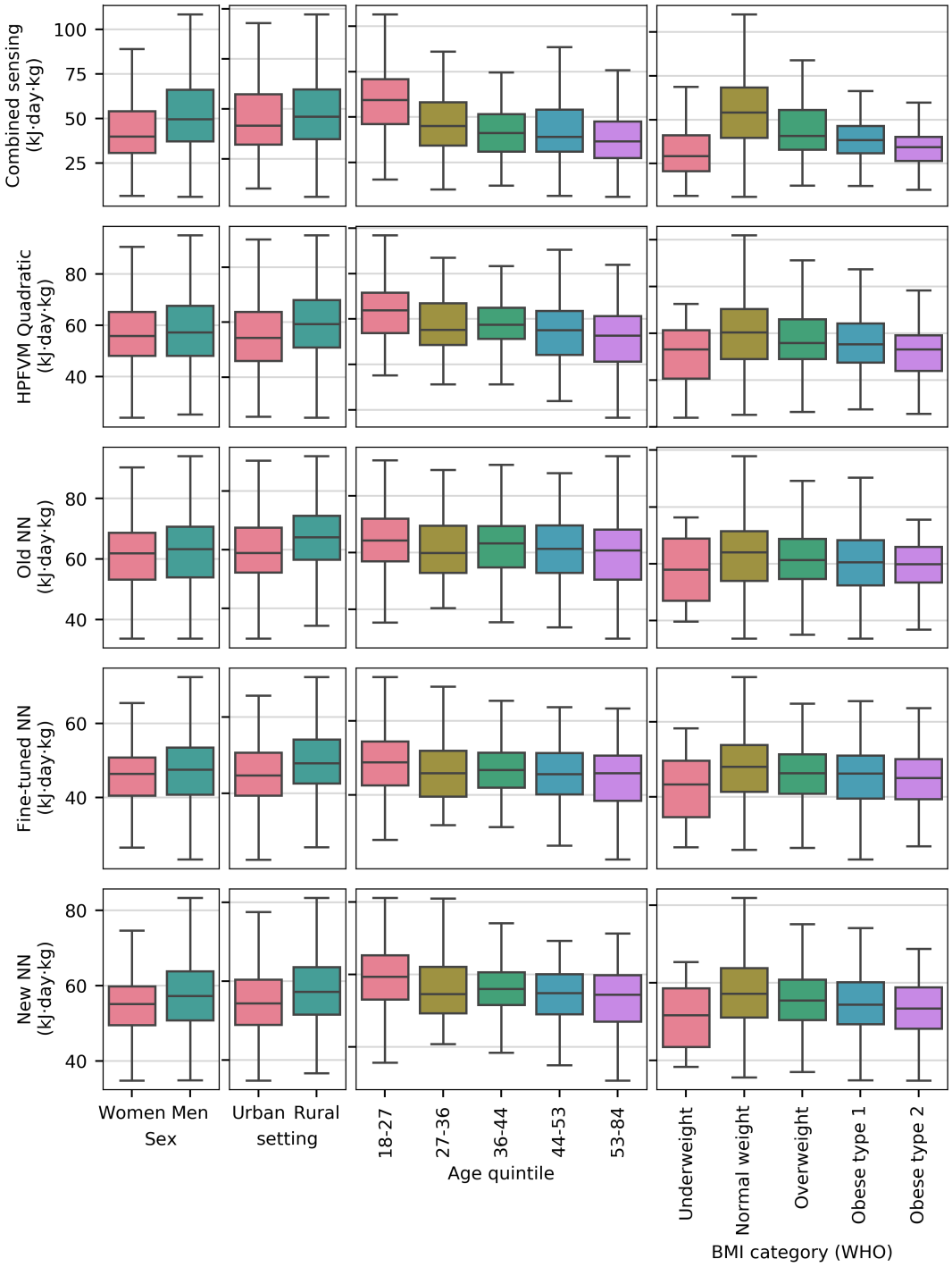


Figure 5.3: Boxplots showing the distribution of estimated activity energy expenditure by each estimation model, across various relevant population strata.

| | Age | Sex | BMI | Urban | Intercept | $r^2$ |
|---|---|---|---|---|---|---|
| Combined sensing | -0.022 | 0.274 | -0.051 | -0.223 | 2.255 | 0.247 |
| HPFVM quadratic | -0.020 | 0.008* | -0.019 | -0.387 | 1.519 | 0.111 |
| Original neural network | -0.011 | 0.065* | -0.015 | -0.435 | 1.045 | 0.073 |
| Fine-tuned neural network | -0.013 | 0.056* | -0.019 | -0.412 | 1.206 | 0.080 |
| New neural network | -0.016 | 0.172 | -0.020 | -0.373 | 1.319 | 0.103 |
| Trunk acceleration | -0.022 | 0.500 | -0.023 | -0.109* | 1.358 | 0.203 |
| Wrist acceleration (HPFVM) | -0.020 | 0.013* | -0.019 | -0.383 | 1.511 | 0.110 |
| Wrist acceleration (ENMO) | -0.021 | -0.053* | -0.019 | -0.353 | 1.579 | 0.112 |

Table 5.5: Coefficients describing the joint relationships between estimated activity energy expenditure and age, sex, BMI, and urban dwelling. An asterisk (*) following a coefficient indicates its 95% confidence interval overlapped 0, suggesting a non-significant association.

Table 5.6 shows the population-level adjusted means of activity energy expenditure, wrist acceleration, and trunk acceleration for the Cameroon 2 study, when matched to the age and BMI means of the Fenland population (age=49.90, BMI=25.68 for women, and age=51.58, BMI=26.99 for men), in which the original movement intensity models were derived [91]. The Fenland population expended significantly more energy per kilo of body weight, but wrist acceleration levels were very similar between the two groups, especially for the urban sample of the Cameroon 2 study.

| Variable | Study (group) | Men | Women |
|---|---|---|---|
| PAEE (J·min$^{-1}$·kg$^{-1}$) | Fenland | 37.4 | 35.5 |
| PAEE (J·min$^{-1}$·kg$^{-1}$) | Cameroon 2 (Urban) | 28.8 (1.3) | 28.7 (1.2) |
| PAEE (J·min$^{-1}$·kg$^{-1}$) | Cameroon 2 (Rural) | 30.2 (1.3) | 32.5 (0.9) |
| Wrist acceleration (HPFVM, milli-g) | Fenland | 47.9 | 48.0 |
| Wrist acceleration (HPFVM, milli-g) | Cameroon 2 (Urban) | 46.7 (1.3) | 49.1 (1.2) |
| Wrist acceleration (HPFVM, milli-g) | Cameroon 2 (Rural) | 49.8 (1.2) | 54.9 (0.9) |
| Trunk acceleration (m·s$^{-2}$) | Fenland | 0.12 | 0.13 |
| Trunk acceleration (m·s$^{-2}$) | Cameroon 2 (Urban) | 0.11 (0.006) | 0.10 (0.004) |
| Trunk acceleration (m·s$^{-2}$) | Cameroon 2 (Rural) | 0.11 (0.005) | 0.10 (0.003) |

Table 5.6: Adjusted means and standard errors of activity energy expenditure, wrist acceleration, and trunk acceleration for the Cameroon 2 study, when matched to the age and BMI means of the Fenland population (age=49.90, BMI=25.68 for women, and age=51.58, BMI=26.99 for men). Fenland mean values included for reference.

Lastly, the correlations between each of the five estimates of activity energy expenditure are shown as a heatmap in Figure 5.4. All of the wrist-based estimates were very highly correlated, from r=0.92 to r=0.98.

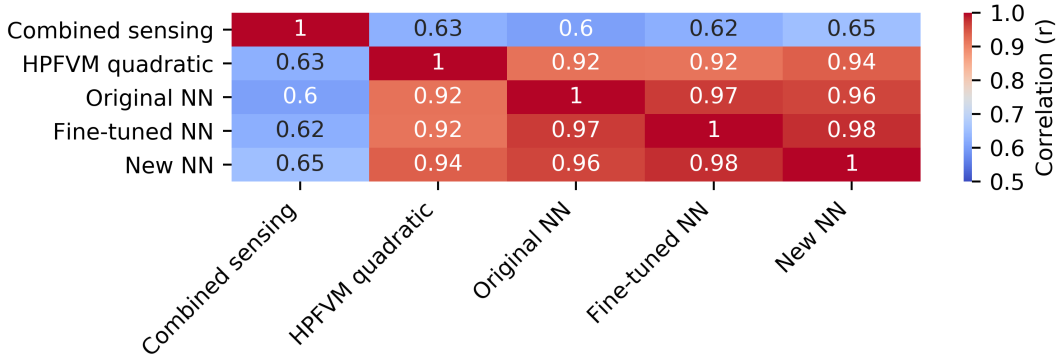


Figure 5.4: Heatmap showing the correlations between each of the five activity energy expenditure estimates, across all participants.

## 5.4 Discussion

In this study, we have derived two deep neural networks to estimate activity energy expenditure from wrist acceleration records collected in free-living Cameroonian adults, and together with two pre-existing models derived in British adults, assessed their agreement with individually-calibrated combined sensing estimates. Our results indicate that the relationship between wrist acceleration and activity energy expenditure is different in British and Cameroonian adults, leading the models to systemic overestimations of activity at the population level. However, it also appears that regardless of the estimation model, similar patterns of activity energy expenditure can be observed with respect to characteristics such as BMI and age.

In our regressions performed within the Cameroonian data to examine the relationship between movement intensity and population characteristics, we observed interesting differences between wrist acceleration and trunk acceleration (Table 5.5). The results indicate that independently of the other factors included, trunk acceleration was higher in men, but there were no sex differences in wrist acceleration. Conversely, urban/rural status was not significantly associated with trunk acceleration, but was strongly associated with wrist acceleration. These contrasts might be explained by differences in activity types between those population strata, resulting in the engagement of different muscle groups and ultimately a difference in the recorded movement profiles. A close similarity in wrist movement intensity between the sexes has previously been observed in large scale British studies [91, 25].

The two wrist-based estimation models derived in British adults both overestimated significantly at the population level, but their high correlations with combined sensing (r=0.62) suggest that they are still effective at ranking people by overall activity level. A higher positive bias was observed in women by all wrist-based estimation models,

and the test sample included more women than men. Wrist-based estimates were much more strongly correlated with combined sensing in rural rather than urban participants, which is particularly surprising for the original models which were derived in Western, largely urban participants, whom we might expect to be more similar in the activities they engage in. These results highlight the critical need for validation of estimation models in local populations before deployment, in order to avoid misleading conclusions regarding differences in activity energy expenditure between populations. Furthermore, if our speculation is correct that these estimation errors are driven by systematic differences in activity type composition, other accelerometry based inferences such as activity recognition may be affected, therefore they also may require a local validation strategy.

The adjusted means for the Cameroon 2 study suggest that at an age of approximately 50 years and a BMI of approximately 26.5 kg·m$^{-2}$, the ratio of wrist movement to activity energy expenditure is fundamentally different to that of British adults in the Fenland study [91]. Fenland participants expended significantly more activity energy expenditure relative to their body weight (mean=36 J·min$^{-1}$·kg$^{-1}$, versus 29 and 31 J·min$^{-1}$·kg$^{-1}$ in the urban and rural groups, respectively), whereas adjusted wrist acceleration intensity was very similar (mean=48 milli-g in Fenland, versus 48 and 52 milli-g in the urban and rural groups, respectively). The difference in these population-level values suggests that a simple movement intensity equation (which assumes a linear model) is not likely to accurately capture both relationships.

When we originally presented our neural network methodology for estimating activity energy expenditure, we conjectured that such complex models have a greater potential for population specificity than simpler models (chapter 4). In our agreement analyses, we have found indications of population specificity for *both* the simple and complex models that were originally derived in British adults, in the form of large positive esti-

mation biases (over-estimation), which were even higher for the neural network model. Of the two neural networks derived in the locally-sourced data, only the fine-tuned network was found to have no statistically significant estimation bias at the population level. This occurred despite both models being chosen from their respective training runs by the same selection strategy (by optimal agreement in the shared independent validation dataset). These findings support the hypothesis that model refinement using local data is more effective than learning a new model from scratch, but to make comparisons between countries, it would be desirable to base those comparisons on objectively identical inference schemes. Further research is required in order to determine which aspects of the models are failing when applied in other populations, and under what conditions we may derive a "global" model that is robust to these differences.

While we have performed an agreement analysis between wrist-derived estimates and individually-calibrated combined sensing estimates, it must be emphasised that neither qualifies as a gold-standard measure of activity energy expenditure. We note that the estimations by the combined sensing methodology have been consistently shown to be unbiased in a diverse range of countries and settings, which includes a small study in Cameroon (n=33) [6]. Furthermore, combined sensing was implemented in a previous wave of this study, finding very similar levels of activity in the urban population [5]. In our previous experiments, we have consistently found that estimation performance exceeded expectations when ultimately evaluated using a gold-standard (higher correlations, tighter limits of agreement and lower RMSE) compared to a silver-standard ([89] and chapter 3, chapter 4). We suspect this is a natural consequence of the random error inherent in the combined sensing methodology [14], which would naturally attenuate the agreement with other measures without affecting its mean tendency. There is clearly a critical need for criterion validation studies conducted in these under-studied countries, particularly given the present opportunities to witness populations undergoing significant lifestyle transitions.

The demographic distribution of the Cameroon 2 study is diverse but not nationally representative, so it cannot necessarily be inferred that these validation results are generalisable to the whole country.

In summary, we have performed agreement analyses between estimates of activity energy expenditure between individually-calibrated combined sensing and four different wrist-based estimates in a large cohort of Cameroonian adults, and found that models derived in British adults overestimated at the population-level, but performed well at sorting the population by activity level. However, a neural network trained on British data and fine-tuned in Cameroonian data performed similarly but with a non-significant mean bias at the population level. At the conclusion of this work, there is no evidence that any single model for the estimation of activity energy expenditure from wrist acceleration appears equally valid in both British and Cameroonian adults; consequently, further work is required before population-level comparisons of activity energy expenditure can be made from data collected exclusively by wrist accelerometry.

## 5.5 Supplementary material

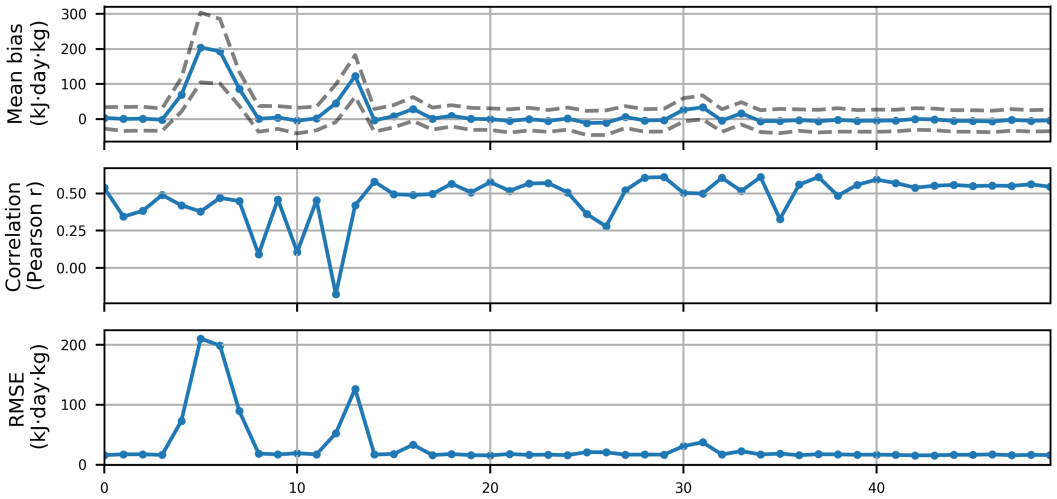


Figure 5.5: Performance of the newly-derived neural network throughout training, evaluated by agreement with individually-calibrated combined sensing in 40 participants.

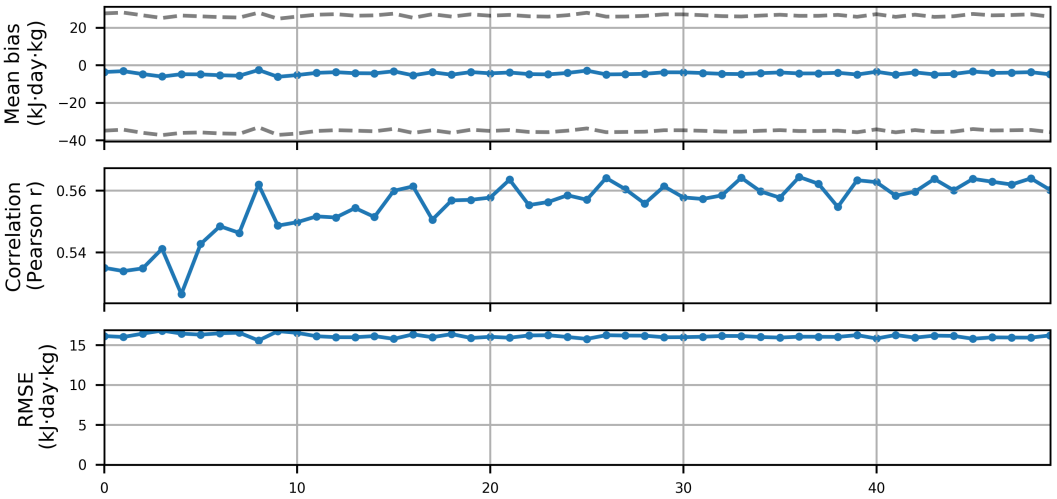


Figure 5.6: Performance of the fine-tuned neural network throughout training, evaluated by agreement with individually-calibrated combined sensing in 40 participants.

Chapter 6

# Comparing models for the estimation of activity energy expenditure using data collected during different activity types

*Introduction: Neural networks have been trained to estimate energy expenditure from raw acceleration data, but the complexity of these models makes them difficult to understand. Insights into the inner workings of such models may help explain how they perform compared to traditional movement intensity approaches, and potentially assist in training better models in future.*

*Methods: 28 adults performed a variety of activities in a laboratory (lying, sitting, standing, walking, and cycling), whilst wearing a triaxial accelerometer on the non-dominant wrist. Activity energy expenditure was estimated during each of these activities using a previously published movement intensity equation and a pre-trained neural network, both derived using data collected during free-living; differences between these estimates was assessed using paired t-tests. The activation levels of neurons in the network were similarly examined for differences by activity type using unpaired t-tests.*

*Results: Across all activities, activity energy expenditure estimates were higher by the neural network than the movement intensity model ($p < 0.0001$), including 13% higher estimates during walking ($p < 0.0001$), and 64% higher during cycling ($p < 0.0001$). There were neurons whose activation was strongly positively or negatively associated with every activity studied; this was still the case after adjusting their activation for movement intensity.*

*Discussion: The difference between the two energy expenditure estimates by activity type highlights that the neural network bases its estimates on more signal features than just movement intensity. This is further supported by the higher activation of neurons during certain activities over others. A neural specificity for patterns corresponding to activity types may be how the model learns to better fit activity energy expenditure in free-living, by implicitly performing activity recognition and adjusting its estimates accordingly.*

## 6.1 Introduction

In many large scale epidemiological studies, accelerometers are being used to capture the physical activity behaviours of participants under free-living conditions [25, 78, 22]. These devices typically measure acceleration in three axes at high frequency, which when attached to a participant results in a high resolution signal describing their movement patterns. As there is a natural relationship between skeletal muscle movement and activity energy expenditure [18], there is a wealth of research aiming to estimate activity energy expenditure from such data [64].

In previous work, two models for the estimation of activity energy expenditure from wrist acceleration have been derived and subsequently validated against doubly labelled water, a gold-standard measure of energy expenditure in free-living [89] (and chapter 3), (chapter 4). The first set of models calculate a movement intensity metric from the raw acceleration signal, which is converted to activity energy expenditure by linear and quadratic equations [91]. The second model is a deep convolutional neural network trained to estimate activity energy expenditure directly from short frames of raw wrist acceleration data (chapter 4). According to the evaluation by agreement with the gold standard measure, both model types were found to outperform previous efforts in this field [64], and the estimates by the neural network were more precise than those of the movement intensity model. When both estimation models were applied in a population of adults in urban and rural Cameroon, there was evidence to suggest that their estimates were positively biased (overestimating activity energy expenditure), and that the neural network model appeared significantly more prone to this overestimation (chapter 5).

Based on a mechanistic understanding of how these two estimation models work, we can hypothesise that they may produce different estimates during different activities. A

movement intensity model estimates higher energy expenditure when the device moves more. The limitations of such an approach have been discussed extensively elsewhere [91]; in brief, different body parts accelerate at different rates during different activities, therefore its accuracy depends on the relative involvement of the measured body part during the activity being performed. For example, wrist movement can be very low during cycling, which can involve high amounts of activity energy expenditure, and would likely be significantly underestimated by a movement intensity model. The rationale behind using a deep neural network is its potential to model highly nonlinear relationships; put simply, it is hoped that a neural network can learn to be robust to these corner-cases, by using patterns in the data to recognise that an activity is being performed where movement intensity alone would give a biased estimate of activity energy expenditure. The evaluations conducted on these models so far cannot confirm if the neural network has learned to make such inferences, nor can they explain under what circumstances we might expect the two estimation models to differ.

The purpose of this study was to apply existing activity energy expenditure estimation models to an annotated dataset where we have certainty on the activities being performed, and to examine the differences between those estimates by activity type. The second purpose was to investigate the internal activations of the neural network in response to different activity stimuli, and ultimately how those responses affect the energy expenditure predictions made by the model.

## 6.2 Methods

The labelled activity data used in this analysis was collected by the Physical Activity Annotation Study (PAAS), the detailed protocol for which can be found elsewhere [81]. Briefly, the study recruited a convenience sample of 28 participants, each of whom per-

formed a set routine of activities whilst wearing 9 GENEA devices [1]. One accelerometer was attached to each of both wrists, both ankles, both hips, the lower back, the upper right arm and the upper right leg; however, only the data collected by the sensor on the non-dominant wrist was included in the following analyses. The devices recorded triaxial acceleration at 80 Hertz with a dynamic range of $\pm$ 6 g. The participants were given prompts to perform each activity via a voice recording, which enabled accurate timestamping of activity start and end times. Prior to the measurement, participants were shown a brief instructional video which demonstrated each activity, to clarify what was expected at each stage.

The activities performed in the PAAS study were designed to be complementary to existing studies by covering more routine everyday activities such as hand washing, and eating with a knife and fork. However, many of these activities were performed for as little as six seconds, which was impractical to examine with the specific neural network model considered here, as it operates on fifteen seconds windows of data at a time. In an effort to preserve as much data and statistical power as possible, we chose to consolidate various labels into one, effectively treating them as the same activity; for example, the participants performed both outdoor cycling on a standard road bike, and indoor cycling on an ergometer, both of which were simply labelled "cycling" for these analyses. Details of which activities were used and consolidated can be found in the supplementary material.

### 6.2.1 Movement intensity and energy expenditure

Two movement intensity models were derived using the non-dominant wrist data, Euclidean Norm Minus One (ENMO) and High Pass Filtered Vector Magnitude (HPFVM). Vector Magnitude (VM) of the raw acceleration signal was calculated as $VM = (X^2 + Y^2 + Z^2)^{0.5}$. From VM, ENMO was calculated as $max(0, VM - 1)$, and HPFVM was cal-

[1] The GENEA was a precursor to the GeneActiv, the device which was used to collect the wrist acceleration data in chapters 2 and 5.

culated by applying a fourth-order Butterworth filter to the VM signal at 0.2 Hertz.

Average activity energy expenditure was estimated for every relevant fifteen second window using two models: 1) a movement intensity quadratic equation was applied to the HPFVM signal according to previously published work [91]: $-1.25+1.1353(HPFVM)-2.4281(\sqrt{HPFVM})-0.0004027(HPFVM^2)$; 2) a deep convolutional neural network (chapter 4) was applied to the raw acceleration data, after it was linearly interpolated to 25 Hertz and reshaped into fifteen-second windows.

The distribution of estimated activity energy expenditure was summarised using means and standard deviations for each activity type, and visualised using boxplots. Differences between the two estimates was assessed within each activity type by paired t-tests. For comparison, the expected range of energy expenditure intensity during the studied activities was retrieved from the Ainsworth compendium [3], a collection of observed energy expenditure levels in a wide range of activities.

### 6.2.2 Neural activations

For each fifteen second non-overlapping time window with a known activity label, the raw acceleration data was fed to the network, and the internal activations of every neuron from the final four layers of the network were calculated. As shown in the model diagram in Figure 4.3, in total this was 14,336 neurons (2,048 from the LSTM layer, and 4,096 in each of the three fully-connected layers). These layers were chosen because the outputs of those neurons are univariate, whereas the preceeding layers output multivariate constructs and would require additional summarisation and inferences. The average activity energy expenditure during that time window was also calculated from the final output layer of the network.

For each neuron, an unpaired t-test was used to test for a difference in its neural activation during each activity compared to every other activity; this yields a t-statistic for each neuron indicating if its activation was higher in walking versus the other four activities

(lying, sitting, standing, cycling), etc. This analysis was also performed for each neuron after pre-residualising its neural activation by movement intensity (specifically HPFVM), yielding a t-statistic indicating if its activation was higher in one activity compared to all the others, over-and-above the activation which might be expected based on movement intensity. Similarly, linear regression was used to characterise the relationship between each neural activation and the average estimated activity energy expenditure.

## 6.3 Results

A summary of the movement intensities and activity energy expenditure predictions for each activity is given in Table 6.1. The distribution of wrist movement intensity during each activity (according to both ENMO and HPFVM metrics) is illustrated in Figure 6.1 using boxplots; wrist movement was highest during walking (mean HPFVM = 186 milli-g), followed distantly by cycling (63 milli-g). Accordingly, the estimates of activity energy expenditure from wrist movement intensity were 162 J·min$^{-1}$·kg$^{-1}$ and 49 J·min$^{-1}$·kg$^{-1}$ for walking and cycling, respectively.

| | | Movement intensity | | Activity energy expenditure | |
|---|---|---|---|---|---|
| | | ENMO | HPFVM | HPFVM quadratic | Neural network |
| | N | (milli-g) | (milli-g) | (J·min$^{-1}$·kg$^{-1}$) | (J·min$^{-1}$·kg$^{-1}$) |
| Lying | 879 | 2.6 (8.0) | 6.5 (14.1) | 2.7 (11.8) | 9.1 (13.0) |
| Sitting | 752 | 13.6 (18.7) | 22.1 (31.8) | 15.0 (27.6) | 17.2 (25.8) |
| Standing | 780 | 33.2 (59.8) | 39.2 (59.0) | 29.5 (52.6) | 39.4 (56.8) |
| Walking | 537 | 173.1 (102.0) | 186.3 (77.7) | 161.8 (69.1) | 182.5 (74.4) |
| Cycling | 723 | 38.1 (36.3) | 62.7 (44.3) | 49.1 (39.7) | 80.3 (54.1) |

Table 6.1: Summary of wrist movement intensity and estimated activity energy expenditure during five activities: lying, sitting, standing, walking, and cycling.

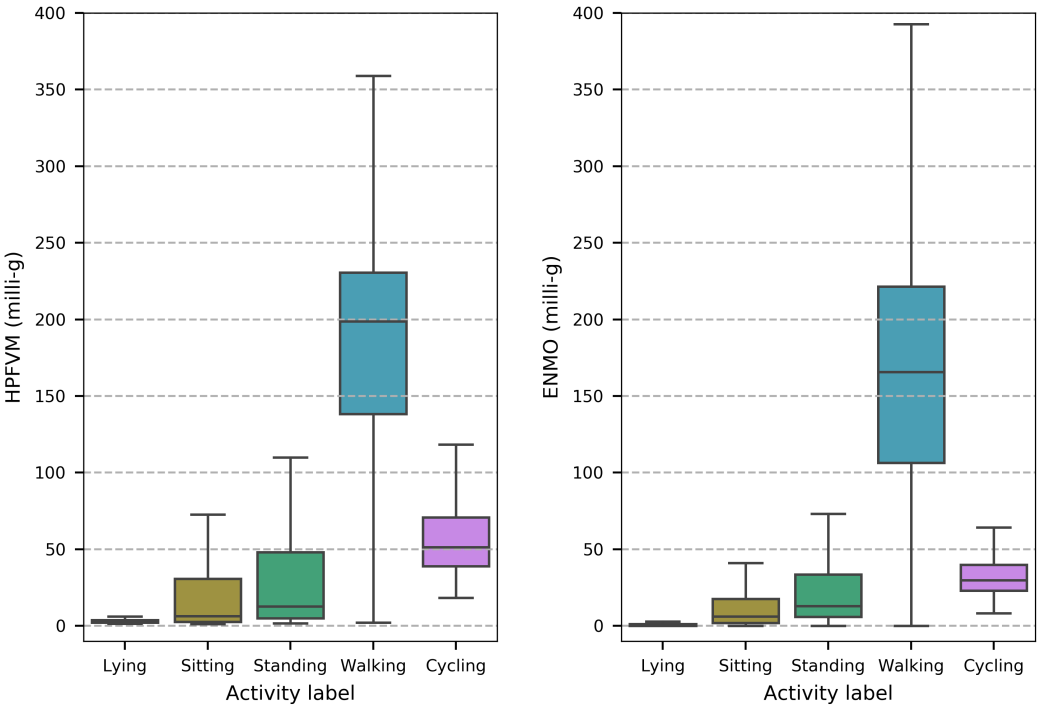


Figure 6.1: Boxplots showing the distribution of wrist movement intensity during lying, sitting, standing, walking, and cycling.

The distributions of estimated activity energy expenditure from the two models, in addition to the reference estimates found in the Ainsworth compendium, are illustrated by boxplots in Figure 6.2. A paired t-test showed that activity energy expenditure estimates by the neural network differed significantly overall from that of the movement intensity model ($p < 9 \times 10^{-147}$). Activity energy expenditure estimates during sitting were not statistically significantly different between models ($p > 0.5$), but very significant differences were observed for every other activity. Mean estimated activity energy expenditure during walking was 13% higher from the neural network ($p < 9 \times 10^{-19}$), and 64% higher during cycling ($p < 9 \times 10^{-118}$).

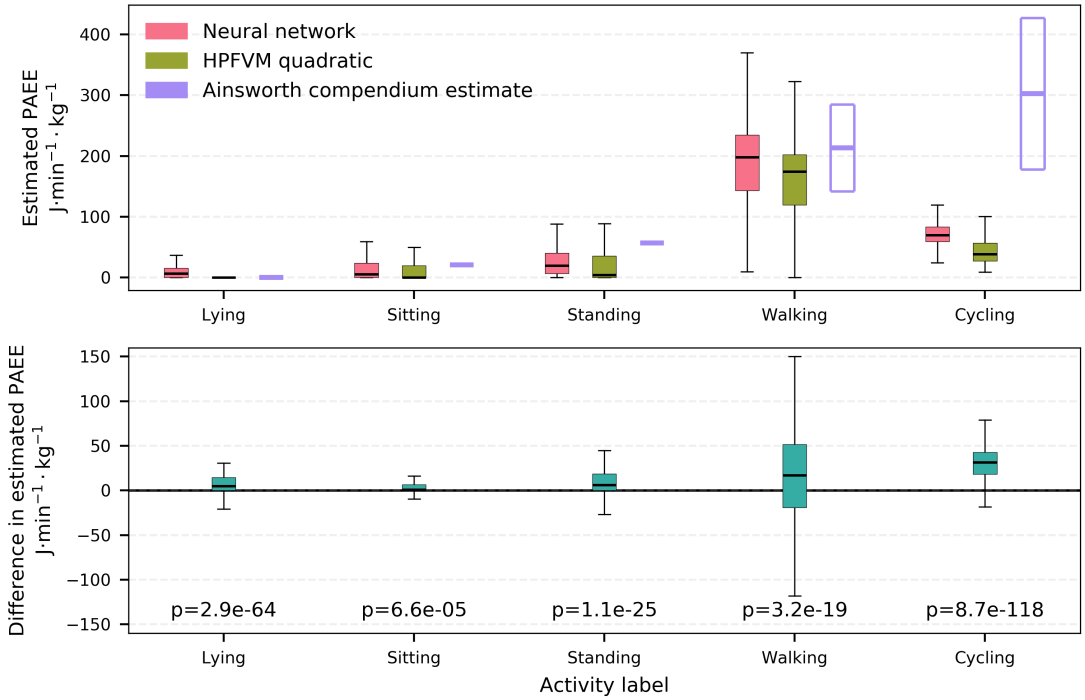


Figure 6.2: Top panel: Boxplots showing the distribution of estimated activity energy expenditure from two different models during lying, sitting, standing, walking, and cycling. Bottom panel: boxplots showing the distribution of pairwise differences between the two models, and the statistical significance of those differences according to paired t-tests.

The reference energy expenditure values retrieved from the Ainsworth compendium[3] are given in Table 6.2. Approximate ranges are given for walking and cycling because the energy cost is highly variable with respect to factors such as speed, terrain, resistance, etc.

| | Reference activity energy expenditure estimates | |
|---|---|---|
| | Net METs | $J \cdot min^{-1} \cdot kg^{-1}$ |
| Lying | 0 | 0 |
| Sitting | 0.3 | 21 |
| Standing | 0.8 | 57 |
| Walking | 2 - 4 | 142 - 285 |
| Cycling | 2.5 - 6 | 178 - 427 |

Table 6.2: Reference activity energy expenditure values from the Ainsworth compendium during five activities: lying, sitting, standing, walking, and cycling.

The associations between neural activity with walking and cycling are visualised in Figure 6.3 using a scatter plot; a dot represents each neuron, the x-axis represents its t-statistic of activation during walking versus all other activities, and the y-axis represents the t-statistic during cycling versus all other activities. Each dot in the top left quadrant therefore represents a neuron whose activation was significantly higher during cycling and significantly lower during walking, whereas dots in the top right represent neurons whose activation was significantly high during walking and cycling versus the other activities. The colour of each neuron represents the direction and strength of its activation association with estimated activity energy expenditure; blue neurons are associated with lower than average estimates, and red neurons are associated with higher than average estimates. The left panel shows these associations without residualisation for movement intensity, and the right panel shows them after residualisation; this shows a significant shift of dots away from “concordant” quadrants (the top right and the bottom left, where the t-statistics are in the same direction) towards “discordant” quadrants (top left and bottom right, where the t-statistics are in opposite directions).

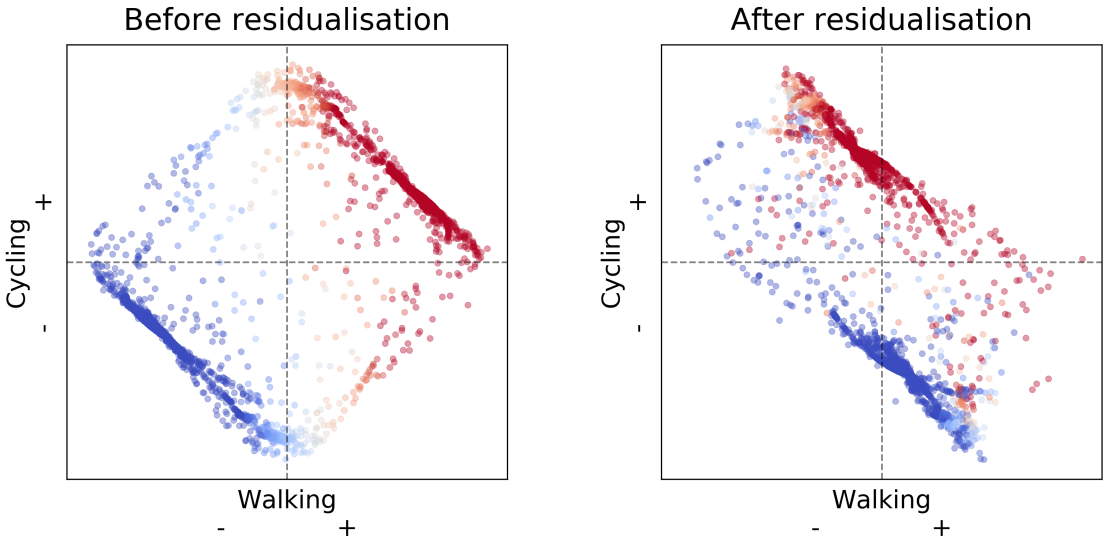


Figure 6.3: Scatter plot showing the association of each neuron's activation during walking versus cycling. Left and right panels show those associations before and after residualisation for movement intensity, respectively.

The visualisation in Figure 6.3 is extended to every activity pair in Figures 6.4 and 6.5, showing the associations before and after adjustment for movement intensity, respectively.

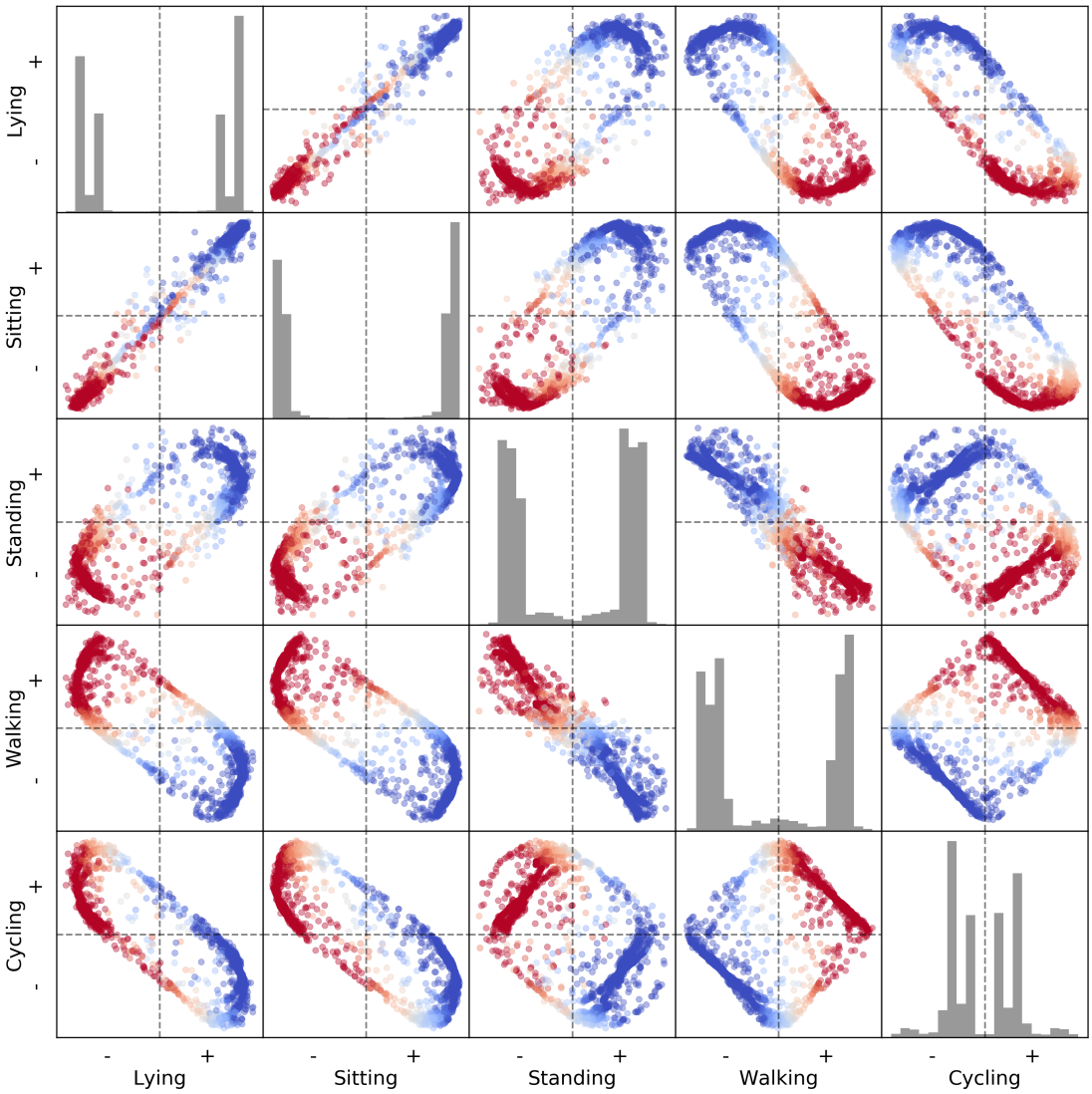


Figure 6.4: Scatter plot showing the association of each neuron's activation during each activity versus every other, before residualisation for movement intensity.

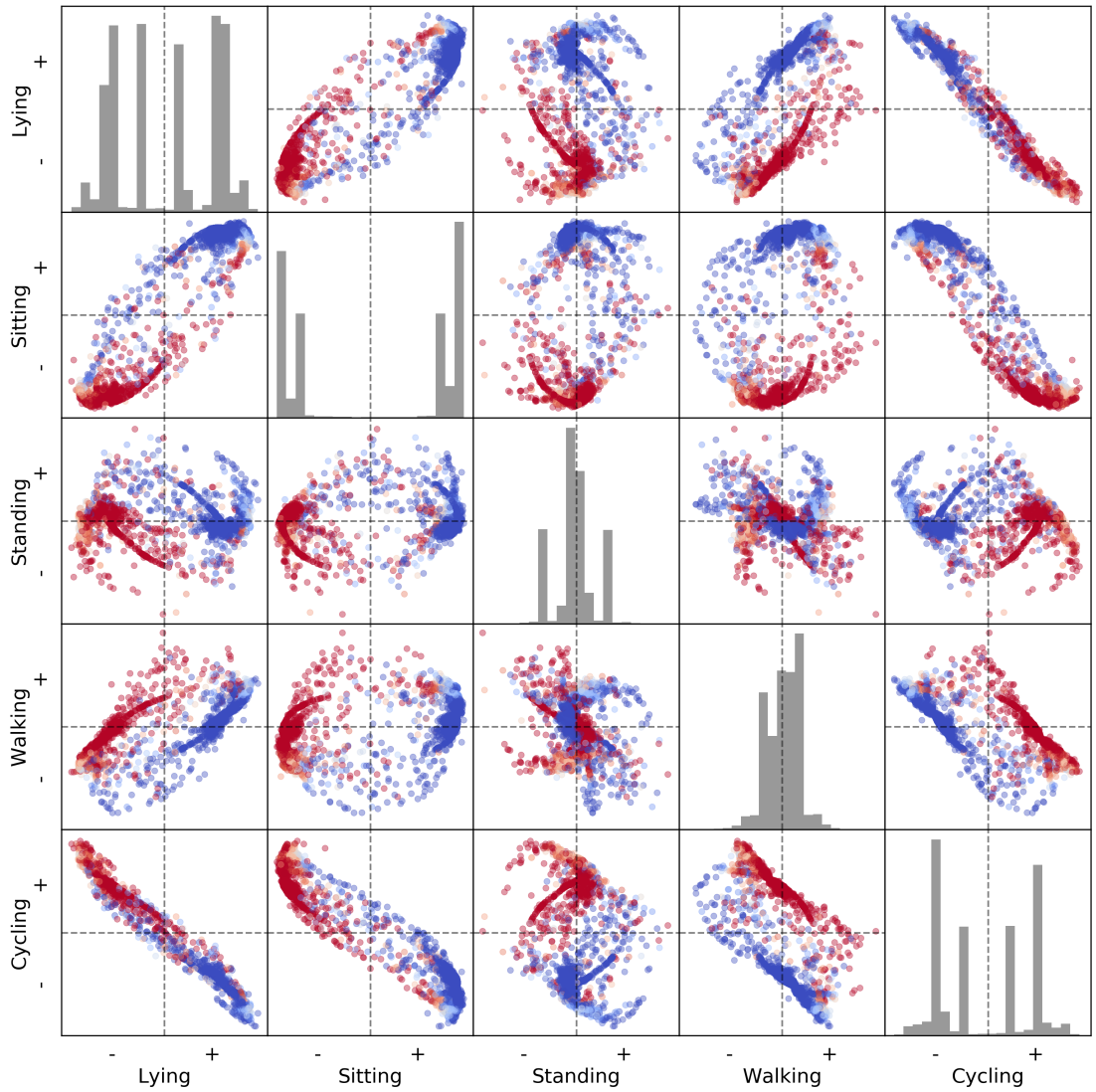


Figure 6.5: Scatter plot showing the association of each neuron's activation during each activity versus every other, after residualisation for movement intensity.

## 6.4 Discussion

In this study, we have used non-dominant wrist acceleration data collected during a number of different routine activities in a laboratory to investigate the differences be-

tween models intended to estimate activity energy expenditure, which were originally derived using data collected during free-living. We observed significant differences in estimated activity energy expenditure based on the activity being performed, most notably during cycling where the neural network estimated 64% higher energy expenditure than the traditional movement intensity model. By examining the activations of the neurons in the final layers of the neural network, we found evidence suggestive of preferential activation for certain activities over others, which give some insight into the mechanism by which the network arrives at estimates which are different from the movement intensity model. These results illustrate the complex nonlinear responses of the neural network, and these insights may lead to a deeper understanding of how estimation performance can vary between individuals and populations.

Estimated activity energy expenditure from both models during all activities except cycling was consistent with previous literature. Comparing against the estimate means from Table 6.1, it can be seen that both models estimate close to these references values for lying, sitting and standing. Estimates from both models during walking were within the given range but towards the lower bound, which is to be expected as most of the activities were deliberately low intensity. The mean estimates during cycling were far below the reference range of between 178 to 427 J·min$^{-1}$·kg$^{-1}$, but the neural network was the closest at 80 J·min$^{-1}$·kg$^{-1}$. The lower estimates from accelerometry may be due to the inclusion of indoor cycling on an ergometer, which was set to automatically adjust its resistance in response to participant cycling speed.

The statistically significant differences in estimated activity energy expenditure between the movement intensity model and the neural network imply that the neural network responds selectively to different sequences and patterns, and is therefore doing something “more intelligent” than simply charactering movement intensity in finer detail. The difference in estimated activity energy expenditure during cycling gives the strongest

evidence in support of this; the results confirm our expectation that wrist movement was low during cycling, and therefore the movement intensity based activity energy expenditure estimates were accordingly low. However, the neural network estimated disproportionally more activity energy expenditure. The dataset upon which the neural network was trained was collected during free-living from adults in Cambridge (United Kingdom), where cycling is a very common mode of travel; it is therefore likely that the training dataset contained many cycling examples, giving the model the opportunity to learn the important differences between (for example) sitting and cycling.

In the analyses of neural activations, there were many neurons whose activation was strongly positively or negatively associated with each activity versus the others, and the histograms of the unadjusted t-values in Figure 6.4 showed clear bimodal distributions of neural activity with respect to each activity type. The adjustment for movement intensity changed these results significantly, indicating that intensity of movement has a large impact on neural activation, but the associations visualised in Figure 6.5 showed that there were still neurons with clear activation preferences for one activity over the others. For example, there were neurons that were positively associated with cycling but negatively associated with walking, even after adjustment for movement intensity. This is significant because these responses are interpreted by subsequent layers of the network and therefore directly contribute to the final energy expenditure estimate, and consequently, it directly explains the mechanism by which the network is capable of estimating higher values during activities like cycling.

The relatively large differences in estimated activity energy expenditure by activity type leads us to speculate that estimation performance must be greatly impacted by the activity type composition of a participant. Furthermore, it may help explain why the neural network overestimated so significantly when applied to data collected in a Cameroonian sample (chapter 5); perhaps there are significant differences at the population level in

activity type composition. There may be an activity which is more common in Cameroon than Britain, whose wrist acceleration signature incorrectly stimulates a disproportionately high energy expenditure prediction, leading to frequent overestimates. If this hypothesis is correct, there are two important implications for extending this methodology to a global surveillance strategy. Firstly, a greater emphasis must be placed on collecting a more varied dataset for model training, to maximise the range of movements the model is exposed to. This may be achieved by supplementing datasets collected during free-living with data collected in a laboratory, which would guarantee the inclusion of certain activities. Secondly, prior to deployment in unfamiliar populations, it would be prudent to verify local face validity in locally-relevant activities, for which the design of the present study can serve as a prototype.

Deep neural networks generally require large amounts of training data in order to converge towards a solution in the performance of complex tasks, and it has been asserted that a lack of such labelled training data is holding back progress in the field of human activity recognition [65]. We have shown in previous work that it is feasible to collect relatively large quantities of simultaneous activity energy expenditure and raw acceleration signals in free-living [91, 89], and those datasets were used to derive the estimation models studied here. If, by training a neural network to solve the energy regression problem, we have implicitly trained it to extract features which distinguish between different activities, this offers a potential solution to the data scarcity problem. It might be advantageous to use this methodology as a model pre-training strategy; first training a network on large amounts of data to do energy expenditure predictions, then reconfiguring the final layers of the network to do activity type predictions, and resuming training using the smaller dataset. A neural network pre-trained in this manner may require less labelled training data.

In conclusion, we have used non-dominant wrist data collected in laboratory conditions

to investigate differences between two existing models designed to estimate activity energy expenditure, and found that they differed significantly by activity type. Estimations were consistently very low for the three low-movement activities (lying, sitting, and standing) from both models, but the neural network estimated higher during walking and significantly higher during cycling. Activations of neurons within the network also differed strongly by activity type. These results provide evidence that deep neural networks have the potential to perform better in scenarios known to be problematic for traditional movement intensity models, by responding to more complex signal features which are indicative of the activity type being performed.

## 6.5 Acknowledgements

Data collection in the PAAS study was conducted by Vincent van Hees.

## 6.6 Supplementary material

### 6.6.1 Activity consolidation

In this study, several variations of the same general activity were performed, but for the purpose of these analyses they were consolidated into just five different activities. The following lists which of the original study labels were consolidated into the overall categories of lying, sitting, standing, walking, and cycling.

Lying:

- inactivity - lying

Sitting:

- activities in the home - sitting + read newspaper
- activities in the home - take a seat and relax
- inactivity - sitting + having conversation
- office part b - sitting
- office part b - sitting on office chair

Standing:

- office part b - standing
- inactivity - standing + having conversation
- indoor walking part - standing

Walking:

- indoor walking part - fast walking
- indoor walking part - slow walking
- indoor walking part - walking
- shopping/street life - walking + carrying bag
- activities in the home - walking to lounge area

Cycling:

- cycling part 1
- cycling part 2
- cycling moderate 1
- cycling moderate 2
- cycling slow 1

- cycling slow 2

# Chapter 7

# Conclusion

## 7.1 Overview

This body of work encompasses a number of methodological advances that expand our ability to interpret triaxial acceleration data collected using body-worn sensors under free-living conditions. The primary goal of this work has been to enable the estimation of activity energy expenditure from such data, and to evaluate the performance of those estimation models in ways relevant to their potential application in a global context. What follows is a brief summary of the contents of each chapter, and a narrative describing the connections between them.

### 7.1.1 Chapter Two

This work started by pursuing the "traditional" approach of relating movement intensity measured at the non-dominant wrist to activity energy expenditure, using linear and quadratic regression models. In previous efforts, the source data was collected during activities in an exercise laboratory [42] or using very few datapoints from summary-level data with a gold-standard measured in free-living [83]. Here the data was collected under free-living conditions using individually-calibrated combined sensing, which is a less precise energy expenditure criterion than respired gas analysis, but allowed the collec-

tion of higher resolution signals in a very large sample, and in a context appropriate to the intended application.

The results revealed that there is a strong linear and curvilinear relationship between wrist movement intensity and activity energy expenditure in free-living UK adults. It was shown that the (cross-sectional) dose-response relationship between estimated activity energy expenditure and BMI was very similar in a large independent sample, whether that estimate came from wrist acceleration or individually-calibrated combined sensing. This demonstrated the epidemiological utility of wrist accelerometry, as it implied that it can be used to accurately describe activity energy expenditure in UK adults, and to observe important aetiological relationships.

### 7.1.2 Chapter Three

The silver-standard criterion used in Chapter Two was insufficient to make conclusive statements about model validity, as it could not be ruled out that we were quantifying correlated error of two lower quality measurements. In Chapter Three, a new dataset was introduced in which simultaneous measurements of both wrists and thigh were collected in free-living conditions, and a gold-standard isotopic measure of total energy expenditure in a subsample. Very high correlations were found between movement intensities at all three anatomical sites, and it was therefore possible to derive linear harmonisation equations to map between them. By combining these models with the activity energy expenditure estimation models from the preceeding chapter, a series of two-step inference models were created to estimate activity energy expenditure from thigh and dominant wrist acceleration intensity.

Finally, these estimation models were applied in the independent sample wherein total energy expenditure was measured by doubly labelled water, and their absolute validity was demonstrated by assessing agreement with activity energy expenditure derived from total energy expenditure and measured resting metabolic rate. It was shown that

agreement with the gold-standard measure was very strong, and nearly all models achieved non-significant estimation biases at the population level, which has never previously been demonstrated from accelerometry measures alone [64].

### 7.1.3 Chapter Four

Using the dataset introduced in the previous chapter, deep convolutional neural networks were trained to predict energy expenditure by directly interpreting raw acceleration data itself, rather than the derived intensity signals used previously. Variations on this technique have been used previously to classify activity types from wearable sensor data; it was hypothesised that their ability to exploit acceleration patterns and model non-linear functions would also be relevant for inferring energy expenditure.

Following the same performance evaluation routine as before against the gold-standard measure of energy expenditure in a separate sample, these estimation models slightly outperformed any other measurement and inference model combination used to estimate activity energy expenditure in free-living adults, including their movement intensity counterparts from Chapter Three. Remarkably, the estimates even appeared to slightly surpass individually-calibrated combined sensing [14], against which those new models were derived.

### 7.1.4 Chapter Five

All of the aforementioned estimation models (relevant to the non-dominant wrist) were applied to a newly-introduced dataset collected in urban and rural Cameroon, which contained simultaneous individually-calibrated combined sensing and non-dominant wrist acceleration. In parallel, a relatively small sample of Cameroonian data was held back to derive two new neural network models; one by fine-tuning the best model from Chapter Three, and one trained from scratch using the same dataset. All wrist-based estimates were then compared against the combined sensing estimates.

The results indicated that all models trained in British data significantly overestimated activity energy expenditure in Cameroonians, but the correlations suggested that they still accurately ranked the population from most to least active. The only model to achieve a non-significant mean bias at the population level was the neural network first trained on British data, then fine-tuned in Cameroonian data. While this agreement analysis was performed against a silver-standard reference, it still suggests that population specificity is a significant concern for both simple and complex energy expenditure models.

### 7.1.5 Chapter Six

The apparent "black-box" nature of deep neural networks may be a barrier to their adoption in practice, as researchers may be concerned about their behaviour when presented with new data. The poorer performance of neural networks in Cameroonian data in Chapter Five highlighted the pressing need to understand trained models, and what about those models can be subject to failure. In this final chapter, wrist acceleration data collected during common activities in a laboratory was used to investigate differences in activity energy expenditure estimation between a movement intensity model derived in Chapter Two, and the neural network derived in Chapter Four. The internal activations of neurons in the neural network were examined for evidence of activity type specificity.

Differences between the two models were observed by activity type, as were differences in activations of neurons within the network. The most striking difference was found in the estimates during cycling; the estimate based on wrist movement alone was predictably very low, and the neural network estimated disproportionately higher energy expenditure. These results suggest that deep neural networks can be robust to prototypical scenarios where traditional movement intensity models are expected to fail. This may also make them more sensitive to more exaggerated estimation fail-

ures during unfamiliar activities, which could explain their shortcomings when applied in Cameroon.

## 7.2 Future work

### 7.2.1 Movement intensity models

The traditional movement intensity model to estimate activity energy expenditure from non-dominant wrist acceleration, derived in Chapter Two using data collected in British adults, was suspected to produce overestimates in Cameroonian adults in Chapter Five. A post-hoc analysis suggested that when matched on age, sex and BMI characteristics, the Cameroonian population moved their wrists more, but expended less energy in free-living. This means that the ratio of movement to energy expenditure was significantly different between the two populations. The fundamental assumption of a movement intensity model is that this ratio is largely constant - if this assumption is not true, it follows that no such model can universally capture the relationship accurately.

One possible solution is to derive a new set of movement intensity equations to fit a more diverse dataset pooled from several countries. However, if our previous observations hold true and the slope of the relationships can be expected to vary systematically between populations, then the likely product is a “compromise” model that performs equally poorly across all populations, but perhaps without estimation bias by population. Alternatively, it might also be suggested that it is acceptable to derive multiple movement intensity models, one for each population. One challenge for such an approach is that prior to application in a new population where no model currently exists, it will need to be determined which model is the most appropriate, a decision which would be difficult to justify in a clear and objective manner.

### 7.2.2 Neural network models

The collective results of this thesis suggest that deep neural networks have a greater potential to model the relationship between raw acceleration data and activity energy expenditure than acceleration magnitude alone. When the neural network models were derived in Chapter Four, it was shown that their estimates agreed slightly more strongly with a gold standard criterion. In Chapter Six, it was shown that they are capable of more complex modelling of activities such as cycling, which may be the source of their performance advantage. However, it was highlighted in Chapter Five that generalisability of such models to other populations may be a significant challenge, perhaps more so than with the movement intensity models. With this challenge in mind, what follows is a brief summary of directions in which this work can be developed further.

#### Model architectures

For brevity, only one neural network architecture was considered in Chapters Four and Five. As noted in the discussion of Chapter Four, the network was one of a virtually infinite number of possible topologies, if we overlook current hardware restrictions. A more thorough and systematic exploration of model breadth and depth would be informative; it is likely that a better performing network exists, but it is impossible to speculate on what performance gains can be expected.

#### Ensembles

In this thesis, only standalone neural networks have been discussed and evaluated. It is common practice to train several models on independent datasets, and to combine together their estimations in the hope of achieving a more robust “ensemble” estimate. An ensemble helps protect against estimation outliers, and works best when estimation errors of the contained models are uncorrelated [23]. An international ensemble could

already be constructed using the networks trained using the British and Cameroonian datasets.

### Unsupervised pre-training

The neural network models which were derived in Chapters Four and Five were trained in a "supervised" fashion; given a dataset of designated inputs and outputs, learn to map the inputs to the outputs. One plausible explanation for the population specificity described in Chapter Five is that the supervised approach is greedy, such that it (understandably) focuses only on exploiting the specific features it witnesses in the training data. In the domain of acceleration data collected by a wearable sensor, this means the network has learned to be familiar with signals corresponding to activities such as cycling, and can understand where an input observation belongs in the distribution of possible inputs. However, when presented with an unfamiliar signal, the observation lies outside of its expected distribution, leading to less stable estimates.

This problem may be remedied by first using an unsupervised learning approach, such as training autoencoder neural networks. Autoencoders, such as variational autoencoders [48] or denoising autoencoders [85], are neural networks which try to learn representations of data. These architectures are thought of as containing two complementary parts: an "encoder" which attempts to transform the input data into some representation, and the "decoder" which attempts to reconstruct the input data from this representation. The network is trained to minimise the reconstruction error of this entire process over a dataset of inputs alone. To optimally perform this task, the network has to capture the whole distribution of possible inputs, by inventing a latent high-dimensional space into which it can map any possible observation.

When training is complete, the "decoder" half of the network can be discarded. The trained "encoder" could then be repurposed as a feature extractor, and the overall task

can then be reformulated as estimating activity energy expenditure from encoded representations. The advantage of this approach is that it can utilise the large quantities of unlabelled data already collected around the world, and a more complete distribution can be defined using greater volumes of data. This pre-training approach has been proposed as a possible use of the UK Biobank accelerometry sub-study [25], as part of a larger effort to advance human activity recognition [65].

### 7.2.3 Future data collections

Datasets for model derivation

The evidence collected together in this thesis suggests that with current methodologies, it is difficult to derive an activity energy expenditure estimation model which can be expected to perform equally well in another population. The methodological suggestions detailed previously may assist in training more robust neural networks which generalise better, but there is likely no perfect substitute for pooling together a more heterogeneous dataset from multiple countries. This would be particularly beneficial if pursuing the unsupervised pre-training approach, because the encoder will have the opportunity to build an encoding scheme that encompasses international variation.

The datasets used for model training in previous chapters were all collected in free-living conditions. A disadvantage of this collection strategy is that most free-living participants spend very little time in high intensity physical activity, which results in a skew of the data distribution towards zero activity. It is also likely that approximately $1/3^{rd}$ of the data describes sleep, assuming an average sleep duration of 7-8 hours per day. This distribution of data is representative of daily life (by definition), but it still means that model training is focused mostly on the ability to distinguish between low and very low energy expenditure values. Selectively filtering the training dataset to contain disproportionately fewer low-intensity observations may balance these priorities during training, potentially leading to more precise estimates in the higher intensity range. Alternatively

or additionally, datasets collected in free-living could be supplemented with exercise data collected in a laboratory, which has the added advantage that under those conditions, the energy expenditure label can be collected with greater accuracy.

### Datasets for model evaluation

Historically, studies designed to validate activity energy expenditure estimation models have been relatively small (typically less than 30 participants) and confined to specific populations [64], due to the extreme costs of criterion measures such as doubly labelled water. The results in Chapter Five, demonstrating an apparent lack of transferability between populations, suggest that a larger scale effort will be required to establish validity at an international scale. An ideal dataset would involve every country in the world periodically measuring a nationally-representative sample, with an experimental setup similar to the study described in Chapter Three. Such a study would currently be prohibitively expensive due to the aforementioned costs, but would nonetheless provide a comprehensive solution to global validation.

As a compromise, it may be more feasible to collect adequate quantities of silver-standard data, such as the Cameroonian dataset described in Chapter Five. A variety of validation studies have demonstrated that the individually-calibrated combined sensing methodology (interpreting a heart rate and movement signal using an exercise test for calibration) has produced activity energy expenditure estimates that were not significantly biased at the population level [84, 6, 11, 14]. In short, combined sensing seems to approximate the mean average activity of a population accurately. By extension, population-level agreement of accelerometry-based estimates with combined sensing estimates would be indicative of validity in that population.

Datasets collected in laboratory conditions should also be considered for the future evaluation of energy expenditure models, even if the intended application is assessment

in free-living. The results in Chapter Six were informative, as despite not having a criterion against which to compare, it inspires confidence in estimation models when predictions are within expected ranges for very common activities such as walking. In general, researchers seeking to use activity energy expenditure inference models in new countries or populations may benefit from testing models during activities specific to their locality; poor validity in a common local activity could be used to justify the need for further model training, and the same data could later be used to confirm that the further training resulted in an improvement.

# References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Leandra Abarca-Gómez, Ziad A Abdeen, Zargar Abdul Hamid, Niveen M Abu-Rmeileh, Benjamin Acosta-Cazares, Cecilia Acuin, Robert J Adams, Wichai Aekplakorn, Kaosar Afsana, Carlos A Aguilar-Salinas, et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. *The Lancet*, 390(10113):2627–2642, 2017.

[3] Barbara E Ainsworth, William L Haskell, Melicia C Whitt, Melinda L Irwin, Ann M Swartz, Scott J Strath, William L O Brien, David R Bassett, Kathryn H Schmitz, Patricia O Emplaincourt, et al. Compendium of physical activities: an update of activity codes and met intensities. *Medicine and Science in Sports and Exercise*, 32:498–504, 2000.

[4] Lars Bo Andersen, Maarike Harro, Luis B Sardinha, Karsten Froberg, Ulf Ekelund, Søren Brage, and Sigmund Alfred Anderssen. Physical activity and clustered cardiovascular risk in children: a cross-sectional study (the european youth heart

study). *The Lancet*, 368(9532):299–304, 2006.

[5] Felix K Assah, Ulf Ekelund, Søren Brage, Jean Claude Mbanya, and Nicholas J Wareham. Urbanization, physical activity, and metabolic health in sub-saharan africa. *Diabetes Care*, 34(2):491–496, 2011.

[6] Felix K Assah, Ulf Ekelund, Søren Brage, Antony Wright, Jean Claude Mbanya, and Nicholas J Wareham. Accuracy and validity of a combined heart rate and motion sensor for the measurement of free-living physical activity energy expenditure in adults in cameroon. *International Journal of Epidemiology*, 2010.

[7] Ling Bao and Stephen S Intille. Activity recognition from user-annotated acceleration data. In *International Conference on Pervasive Computing*, pages 1–17. Springer Berlin Heidelberg, 2004.

[8] K Berkemeyer, K Wijndaele, T White, AJM Cooper, R Luben, K Westgate, SJ Griffin, KT Khaw, NJ Wareham, and Søren Brage. The descriptive epidemiology of accelerometer-measured physical activity in older adults. *International Journal of Behavioral Nutrition and Physical Activity*, 13(1):1, 2016.

[9] Sheila A Bingham, Caroline Gill, Ailsa Welch, Aedin Cassidy, Shirley A Runswick, Suzy Oakes, Robert Lubin, David I Thurnham, TJ Key, Lynn Roe, et al. Validation of dietary assessment methods in the uk arm of epic using weighed records, and 24-hour urinary nitrogen and potassium and serum vitamin c and carotenoids as biomarkers. *International Journal of Epidemiology*, 26(suppl 1):S137, 1997.

[10] UK Biobank. Category 2 enhanced phenotyping at baseline assessment visit in last 100–150,000 participants. =http://www.ukbiobank.ac.uk/wp-content/uploads/2011/06/Protocol_addendum_2.pdf.

[11] Søren Brage, Niels Brage, Paul W Franks, Ulf Ekelund, and Nicholas J Wareham. Reliability and validity of the combined heart rate and movement sensor actiheart.

*European Journal of Clinical Nutrition*, 59(4):561–570, 2005.

[12] Søren Brage, Niels Brage, Paul W Franks, Ulf Ekelund, Man-Yu Wong, Lars Bo Andersen, Karsten Froberg, and Nicholas J Wareham. Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. *Journal of Applied Physiology*, 96(1):343–351, 2004.

[13] Søren Brage, Ulf Ekelund, Niels Brage, Mark A Hennings, Karsten Froberg, Paul W Franks, and Nicholas J Wareham. Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity. *Journal of Applied Physiology*, 103(2):682–692, 2007.

[14] Søren Brage, Kate Westgate, Paul W Franks, Oliver Stegle, Antony Wright, Ulf Ekelund, and Nicholas J Wareham. Estimation of free-living energy expenditure by heart rate and movement sensing: A doubly-labelled water study. *PLoS One*, 10(9):e0137206, 2015.

[15] Søren Brage, Kate Westgate, Katrien Wijndaele, Job Godinho, Simon Griffin, and Nick Wareham. Evaluation of a method for minimising diurnal information bias in objective sensor data. In *International Conference of Ambulatory Monitoring and Physical Activity Measurement*, 2013.

[16] Wendy J Brown, Adrian E Bauman, Fiona C Bull, and Nicola W Burton. Development of evidence-based physical activity recommendations for adults (18-64 years): report prepared for the australian government department of health, august 2012. 2013.

[17] LRSM Cart. Letter to the editor: standardized use of the terms “sedentary” and “sedentary behaviours”. *Appl. Physiol. Nutr. Metab.*, 37(3):540, 2012.

[18] Carl J Caspersen, Kenneth E Powell, and Gregory M Christenson. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Reports*, 100(2):126, 1985.

[19] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042, 2013.

[20] François Chollet et al. Keras. `https://keras.io`, 2015.

[21] Harmon Craig. Isotopic standards for carbon and oxygen and correction factors for mass-spectrometric analysis of carbon dioxide. *Geochimica et cosmochimica acta*, 12(1-2):133–149, 1957.

[22] Inácio CM da Silva, Vincent T van Hees, Virgílio V Ramires, Alan G Knuth, Renata M Bielemann, Ulf Ekelund, Søren Brage, and Pedro C Hallal. Physical activity levels in three brazilian birth cohorts as assessed with raw triaxial wrist accelerometry. *International Journal of Epidemiology*, 43(6):1959–1968, 2014.

[23] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[24] Olivier Dieu, Jacques Mikulovic, Paul S Fardy, Gilles Bui-Xuan, Laurent Béghin, and Jérémy Vanhelst. Physical activity using wrist-worn accelerometers: comparison of dominant and non-dominant wrist. *Clinical physiology and functional imaging*, 37(5):525–529, 2017.

[25] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm Granat, Tom White, Vincent van Hees, Mike Trenell, Chris Owen, Rob Gillions, Simon Sheard, Tim Peakman, Søren Brage, and Nicholas J Wareham.

Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. *PLoS One*, 2017.

[26] Ulf Ekelund, Jostein Steene-Johannessen, Wendy J Brown, Morten Wang Fagerland, Neville Owen, Kenneth E Powell, Adrian Bauman, I-Min Lee, Lancet Physical Activity Series, Lancet Sedentary Behaviour Working Group, et al. Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? a harmonised meta-analysis of data from more than 1 million men and women. *The Lancet*, 388(10051):1302–1310, 2016.

[27] M Elia and Geoffrey Livesey. Theory and validity of indirect calorimetry during net lipid synthesis. *The American Journal of Clinical Nutrition*, 47(4):591–607, 1988.

[28] Katherine Ellis, Jacqueline Kerr, Suneeta Godbole, Gert Lanckriet, David Wing, and Simon Marshall. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiological measurement*, 35(11):2191, 2014.

[29] Dale W Esliger, Ann V Rowlands, Tina L Hurst, Michael Catt, Peter Murray, and Roger G Eston. Validation of the genea accelerometer. *Medicine and Science in Sports and Exercise*, 2011.

[30] Leopold K Fezeu, Felix K Assah, Beverley Balkau, Dora S Mbanya, André-Pascal Kengne, Paschal K Awah, and Jean-Claude N Mbanya. Ten-year changes in central obesity and bmi in rural and urban cameroon. *Obesity*, 16(5):1144–1147, 2008.

[31] Rajna Golubic, Anne M May, Kristin Benjaminsen Borch, Kim Overvad, Marie-Aline Charles, Maria Jose Tormo Diaz, Pilar Amiano, Domenico Palli, Elisavet Valanou, Matthaeus Vigl, et al. Validity of electronically administered recent physical activity questionnaire (rpaq) in ten european countries. *PloS one*, 9(3):e92829, 2014.

[32] P Margaret Grant, Cormac G Ryan, William W Tigbe, and Malcolm H Granat. The validation of a novel activity monitor in the measurement of posture and motion during everyday activities. *British Journal of Sports Medicine*, 40(12):992–997, 2006.

[33] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):11:1–11:28, 2017.

[34] Pedro C Hallal, Lars Bo Andersen, Fiona C Bull, Regina Guthold, William Haskell, Ulf Ekelund, Lancet Physical Activity Series Working Group, et al. Global physical activity levels: surveillance progress, pitfalls, and prospects. *The Lancet*, 380(9838):247–257, 2012.

[35] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[36] Heather A Haugen, Edward L Melanson, Zung Vu Tran, Jay T Kearney, and James O Hill. Variability of measured resting metabolic rate. *The American journal of clinical nutrition*, 78(6):1141–1145, 2003.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[38] Genevieve N Healy, David W Dunstan, Jo Salmon, Ester Cerin, Jonathan E Shaw, Paul Z Zimmet, and Neville Owen. Breaks in sedentary time beneficial associations with metabolic risk. *Diabetes Care*, 31(4):661–666, 2008.

[39] Genevieve N Healy, Katrien Wijndaele, David W Dunstan, Jonathan E Shaw, Jo Salmon, Paul Z Zimmet, and Neville Owen. Objectively measured seden-

tary time, physical activity, and metabolic risk the australian diabetes, obesity and lifestyle study (ausdiab). *Diabetes Care*, 31(2):369–371, 2008.

[40] CJK Henry. Basal metabolic rate studies in humans: measurement and development of new equations. *Public Health Nutrition*, 8(7a):1133–1152, 2005.

[41] Hans-Olav Hessen and Astrid Johnsen Tessem. Human activity recognition with two body-worn accelerometer sensors. Master's thesis, NTNU, 2016.

[42] Maria Hildebrand, VT Hees VAN, Bjorge Hermann Hansen, and Ulf Ekelund. Age group comparability of raw accelerometer output from wrist-and hip-worn monitors. *Medicine and science in sports and exercise*, 46(9):1816–1824, 2014.

[43] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[44] Eric Jéquier. Pathways to obesity. *International Journal of Obesity*, 26(S2):S12, 2002.

[45] Darcy L Johannsen, Miguel Andres Calabro, Jeanne Stewart, Warren Franke, Jennifer C Rood, and Gregory J Welk. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. *Medicine and science in sports and exercise*, 42(11):2134–2140, 2010.

[46] Tuomas O Kilpeläinen, Lu Qi, Søren Brage, Stephen J Sharp, Emily Sonestedt, Ellen Demerath, Tariq Ahmad, Samia Mora, Marika Kaakinen, Camilla Helene Sandholt, et al. Physical activity attenuates the influence of fto variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS Medicine*, 8(11), 2011.

[47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[48] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[49] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[50] I-Min Lee, Eric J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair, Peter T Katzmarzyk, Lancet Physical Activity Series Working Group, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*, 380(9838):219–229, 2012.

[51] Paul Lukowicz, Holger Junker, and Gerhard Tröster. Automatic calibration of body worn acceleration sensors. In *International Conference on Pervasive Computing*, pages 176–181. Springer, 2004.

[52] Kate Lyden, Sarah Kozey Keadle, John Staudenmayer, and Patty S Freedson. A method to estimate free-living active and sedentary behavior from an accelerometer. *Medicine and Science in Sports and Exercise*, 46(2):386, 2014.

[53] Ralph Maddison, Cliona Ni Mhurchu, Yannan Jiang, Stephen Vander Hoorn, Anthony Rodgers, Carlene MM Lawes, and Elaine Rush. International physical activity questionnaire (ipaq) and new zealand physical activity questionnaire (nzpaq): a doubly labelled water validation. *International Journal of Behavioral Nutrition and Physical Activity*, 4(1):62, 2007.

[54] Amy E Mark and Ian Janssen. Influence of bouts of physical activity on overweight in youth. *American Journal of Preventive Medicine*, 36(5):416–421, 2009.

[55] Charles E Matthews, S Keadle Kozey, Steven C Moore, Dale S Schoeller, Raymond J Carroll, Richard P Troiano, and Joshua N Sampson. Measurement of active and sedentary behavior in context of large epidemiologic studies. *Medicine and science in sports and exercise*, 50(2):266–276, 2018.

[56] AH Montoye, Lanay M Mudd, Subir Biswas, and Karin A Pfeiffer. Energy expenditure prediction using raw accelerometer data in simulated free living. *Medicine and science in sports and exercise*, 47(8):1735–1746, 2015.

[57] Alexander HK Montoye, Munni Begum, Zachary Henning, and Karin A Pfeiffer. Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. *Physiological measurement*, 38(2):343, 2017.

[58] Alexander HK Montoye, James M Pivarnik, Lanay M Mudd, Subir Biswas, and Karin A Pfeiffer. Wrist-independent energy expenditure prediction models from raw accelerometer data. *Physiological measurement*, 37(10):1770, 2016.

[59] Angela A Mulligan, Robert N Luben, Amit Bhaniani, David J Parry-Smith, Laura O'Connor, Anthony P Khawaja, Nita G Forouhi, Kay-Tee Khaw, Adam Dickinson, Nick Wareham, et al. A new tool for converting food frequency questionnaire data into nutrient and food group values: Feta research methods and availability. *BMJ Open*, 4(3), 2014.

[60] S Nielsen, DD Hensrud, S Romanski, James A Levine, B Burguera, and Michael Dennis Jensen. Body composition and resting energy expenditure in humans: role of fat, fat-free mass and extracellular fluid. *International journal of obesity*, 24(9):1153, 2000.

[61] Laura O'Connor, Søren Brage, Simon J Griffin, Nicholas J Wareham, and Nita G Forouhi. The cross-sectional association between snacking behaviour and measures of adiposity: the fenland study, uk. *British Journal of Nutrition*, 114(08):1286–1293, 2015.

[62] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[63] Mark Orme, Katrien Wijndaele, Stephen J Sharp, Kate Westgate, Ulf Ekelund, and Søren Brage. Combined influence of epoch length, cut-point and bout duration on accelerometry-derived physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 11(1):1, 2014.

[64] G Plasqui, AGb Bonomi, and KR Westerterp. Daily physical activity assessment with accelerometers: new insights and validation studies. *Obesity Reviews*, 14(6):451–462, 2013.

[65] Thomas Plotz and Yu Guan. Deep learning for human activity recognition in mobile computing. *Computer*, (5):50–59, 2018.

[66] Alessandra Prioreschi, Thomas Nappey, Kate Westgate, Patrick Olivier, Soren Brage, and Lisa Kim Micklesfield. Development and feasibility of a wearable infant wrist band for the objective measurement of physical activity using accelerometery. *Pilot and Feasibility Studies*, 4(1):60, 2018.

[67] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*, pages 233–240. IEEE, 2010.

[68] Alex V Rowlands, Thomas Yates, Tim S Olds, Melanie Davies, Kamlesh Khunti, and Charlotte L Edwardson. Sedentary sphere: Wrist-worn accelerometer-brand independent posture classification. *Medicine and Science in Sports and Exercise*, 48(4):748–754, 2016.

[69] Nasim S Sabounchi, Hazhir Rahmandad, and Alice Ammerman. Best-fitting prediction equations for basal metabolic rate: informing obesity interventions in diverse populations. *International Journal of Obesity*, 37(10):1364, 2013.

[70] LB Sardinha and PB Júdice. Usefulness of motion sensors to estimate energy expenditure in children and adults: a narrative review of studies using dlw. *European journal of clinical nutrition*, 71(3):331, 2017.

[71] Dale A Schoeller. Recent advances from application of doubly labeled water to measurement of human energy expenditure. *The Journal of nutrition*, 129(10):1765–1768, 1999.

[72] Dale A Schoeller, Eric Ravussin, Yves Schutz, Kevin J Acheson, Peter Baertschi, and Eric Jequier. Energy expenditure by doubly labeled water: validation in humans and proposed calculation. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 250(5):R823–R830, 1986.

[73] Miranda T Schram, Simone JS Sep, Carla J van der Kallen, Pieter C Dagnelie, Annemarie Koster, Nicolaas Schaper, Ronald MA Henry, and Coen DA Stehouwer. The maastricht study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European journal of epidemiology*, 29(6):439–451, 2014.

[74] Richard JE Skipworth, Guro B Stene, Max Dahele, Paul O Hendry, Alexandra C Small, David Blum, Stein Kaasa, Peter Trottenberg, Lukas Radbruch, Florian Strasser, et al. Patient-focused endpoints in advanced cancer: criterion-based validation of accelerometer-based activity monitoring. *Clinical Nutrition*, 30(6):812–821, 2011.

[75] Oliver Stegle, Sebastian V Fallert, David JC MacKay, and Søren Brage. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151, 2008.

[76] Scott J Strath, Søren Brage, and Ulf Ekelund. Integration of physiological and accelerometer data to improve physical activity assessment. *Medicine and Science*

*in Sports and Exercise*, 37:S563–71, 2005.

[77] Dylan Thompson, Alan M Batterham, Susan Bock, Claire Robson, and Keith Stokes. Assessment of low-to-moderate intensity physical activity thermogenesis in young adults using synchronized heart rate and accelerometry with branched-equation modeling. *The Journal of Nutrition*, 136(4):1037–1042, 2006.

[78] Rick Troiano and James McClain. Objective measures of physical activity, sleep, and strength in us national health and nutrition examination survey (nhanes) 2011–2014. In *Proceedings of the 8th International Conference on Diet and Activity Methods*, 2012.

[79] Julianne D van der Berg, Coen DA Stehouwer, Hans Bosma, Jeroen HPM van der Velde, Paul JB Willems, Hans HCM Savelberg, Miranda T Schram, Simone JS Sep, Carla JH van der Kallen, Ronald MA Henry, et al. Associations of total amount and patterns of sedentary behaviour with type 2 diabetes and the metabolic syndrome: The maastricht study. *Diabetologia*, 59(4):709–718, 2016.

[80] Vincent T van Hees, Zhou Fang, Joss Langford, Felix Assah, Anwar Mohammad, Inacio CM da Silva, Michael I Trenell, Tom White, Nicholas J Wareham, and Søren Brage. Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents. *Journal of Applied Physiology*, 117(7):738–744, 2014.

[81] Vincent T van Hees, Rajna Golubic, Ulf Ekelund, and Søren Brage. Impact of study design on development and evaluation of an activity-type classifier. *Journal of Applied Physiology*, 114(8):1042–1051, 2013.

[82] Vincent T Van Hees, Lukas Gorzelniak, Emmanuel Carlos Dean Leon, Martin Eder, Marcelo Pias, Salman Taherian, Ulf Ekelund, Frida Renström, Paul W Franks, Alexander Horsch, et al. Separating movement and gravity components in

an acceleration signal and implications for the assessment of human daily physical activity. *PLoS One*, 8(4):e61691, 2013.

[83] Vincent T van Hees, Frida Renström, Antony Wright, Anna Gradmark, Michael Catt, Kong Y Chen, Marie Löf, Les Bluck, Jeremy Pomeroy, Nicholas J Wareham, et al. Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. *PLoS One*, 6(7), 2011.

[84] Clement Villars, Audrey Bergouignan, Julien Dugas, Edwina Antoun, Dale Alan Schoeller, Hubert Roth, Anne-Clemence Maingon, Etienne Lefai, Stéphane Blanc, and Chantal Simon. Validity of combining heart rate and uniaxial acceleration to measure free-living physical activity energy expenditure in young men. *Journal of Applied Physiology*, 113(11):1763–1771, 2012.

[85] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[86] Nicholas J Wareham, Rupert W Jakes, Kirsten L Rennie, Jo Mitchell, Susie Hennings, and Nicholas E Day. Validity and repeatability of the epic-norfolk physical activity questionnaire. *International Journal of Epidemiology*, 31(1):168–174, 2002.

[87] LPE Watson, P Raymond-Barker, C Moran, N Schoenmakers, C Mitchell, L Bluck, VK Chatterjee, DB Savage, and PR Murgatroyd. An approach to quantifying abnormalities in energy expenditure and lean mass in metabolic disease. *European journal of clinical nutrition*, 68(2):234, 2014.

[88] Klaas R Westerterp. Diet induced thermogenesis. *Nutrition & metabolism*, 1(1):5, 2004.

[89] Thomas White, Kate Westgate, Stefanie Hollidge, Michelle Venables, Patrick Olivier, Nick Wareham, and Soren Brage. Estimating energy expenditure from wrist and thigh accelerometry in free-living adults: a doubly labelled water study. *bioRxiv*, 2018.

[90] Tom White. Thomite/pampro v0.4.0, March 2018.

[91] Tom White, Kate Westgate, Nicholas J Wareham, and Søren Brage. Estimation of physical activity energy expenditure during free-living from wrist accelerometry in uk adults. *PLoS One*, 2016.

[92] Katrien Wijndaele, Kate Westgate, Samantha K Stephens, Steven N Blair, Fiona C Bull, Sebastien FM Chastin, David W Dunstan, Ulf Ekelund, Dale W Esliger, Patty S Freedson, et al. Utilization and harmonization of adult accelerometry data: review and expert consensus. *Medicine and Science in Sports and Exercise*, 47(10):2129, 2015.